April 2011

Cluster visualization of upregulated HDAC1 in mouse using integration of Treeview and Galaxy

Timothy Daniel Bonci
Worcester Polytechnic Institute

Cluster visualization of upregulated *HDAC1* in mouse using integration of Treeview and Galaxy

A Major Qualifying Project
submitted to the faculty of
Worcester Polytechnic Institute
in partial fulfillment of the requirements
for the degree of Bachelor of Science

by
Timothy Bonci

Date: April 28, 2011

Report submitted to:

Professor Elizabeth Ryder
Worcester Polytechnic Institute

Table of Contents

Abstract

In order to better understand genetic expression changes in experiments, microarray data is computer analyzed by tools such as the freely available Galaxy, which currently lacks microarray visualization. A visualization interface was built into the toolset using Python. It was enabled with selectable clustering algorithms which were used to analyze the RNA from mice infected with AAV to upregulate *HDAC1* or *LacZ*. The *HDAC* enzymes have been shown to play a part in the negative regulation of types of learning. This clustering identifies other genes as likely actors in the chemistry of memory and learning.

Acknowledgements

We would like to thank Viktor Teplyuk for his programming skills and assistance in maintaining Galaxy for UMass Medical School.

Introduction

The completion of the Human Genome Project in 2003 (Collins et al., 2003) was supposed to herald in a new age of biological understanding, now that we had the keys to understand the underlying genetic structure of the human body. Among other things, it promised a cure for every disease, and genetically personalized health care (Economist, 2010). Of course, that has not yet all come to fruition, even years later. With our own biological data laid bare in front of us, the answers to problems like disease are often complex, nuanced, and not reliant on a single gene to be switched on or off. A focus for this project was learning, memory, and memory-related neurological disorders such as Alzheimer's.

One of the current approaches being studied for disease treatment lies in gene expression. Changing the genome of a living human in order to treat a condition therein would be a difficult undertaking, even when the change is only necessary in certain cells where the condition develops. Applying a compound that promotes a gene to be turned "on" or inhibits it to stay "off" is theoretically a simpler endeavor. One of the goals of this project is to learn about the *HDAC1* gene as a possible target for inhibition in treatment of learning and memory deficiency.

Due to the sheer amount of data available when looking at genomic experiments, computers are used to parse, separate, refine, and highlight the significant details. One freely available, open-source program is Galaxy. This online program is a collection of tools and connected databases developed and maintained by Pennsylvania State University. Galaxy allows biologists to download data from freely shared databases (or upload their own, newly created data) and run computer-assisted analysis. Currently, a user with microarray data from a gene expression experiment needs to use a separate program to cluster the genes and visualize the output data. In working with Dr. Juerg Straubhaar at The University of Massachusetts Medical School, we determined the need to build a tool into Galaxy with a primary functional goal of performing gene expression clustering, while providing an interface with compelling data visualization. We also worked with Dr. Mira Jakovcevski at the University of Massachusetts Medical School, who was interested in using such a tool to analyze the microarray data from her experiments with mice.

Dr. Jakovcevski injected mice with AAV (adeno-associated virus) into their cingulate cortex. The four experimental mice received vectors that upregulated *HDAC1*. The four control mice received upregulated *LacZ*. Dr. Jakovcevski 's experiment is designed to investigate the impairment of functions in the forebrain, such as general drive and working memory, so the upregulation of HDAC1 was localized to the cingulate cortex. This experiment was based off of a finding that showed increased HDAC1 expression in the brains of schizophrenia patients, and the most impaired portion is the forebrain (Sharma, et al., 2008). The mice with increased HDAC1 did show reduced activity and impaired working memory. In the current experiment, we focused on HDAC1 in order to learn more about its function and what operational relatives we could find via clustering.

Background

Clustering and Analysis of Gene Expression Data

Gene expression data are computerized for analysis and usually take the form of a table, with rows being genes, and columns being different experimental conditions (or replications to filter noise). Clustering is a way of re-arranging the rows (or columns) to group together genes with similar expression pathways. Take the following table, a generic example showing the expression of four genes in conditions C1-C4:

Table 1.	C 1	C 2	C 3	C 4
Gene 1	2.1	2.3	8	1.1
Gene 2	9.4	9.1	1.5	5.1
Gene 3	4.1	7.3	0.4	3.4
Gene 4	10	9.5	38.1	4.7

Gene 1 is stable for C1 and C2, spikes at C3, and is lower in C4. Gene 2 is stable in C1 and C2, and then drops in C3 and recovers in C4. Gene 3 fluctuates slightly up and down, and Gene 4 changes similarly to Gene 1. It is stable in C1 and C2, spikes in C3, and then falls in C4. This is easier to see in graphical form (Figure 1).

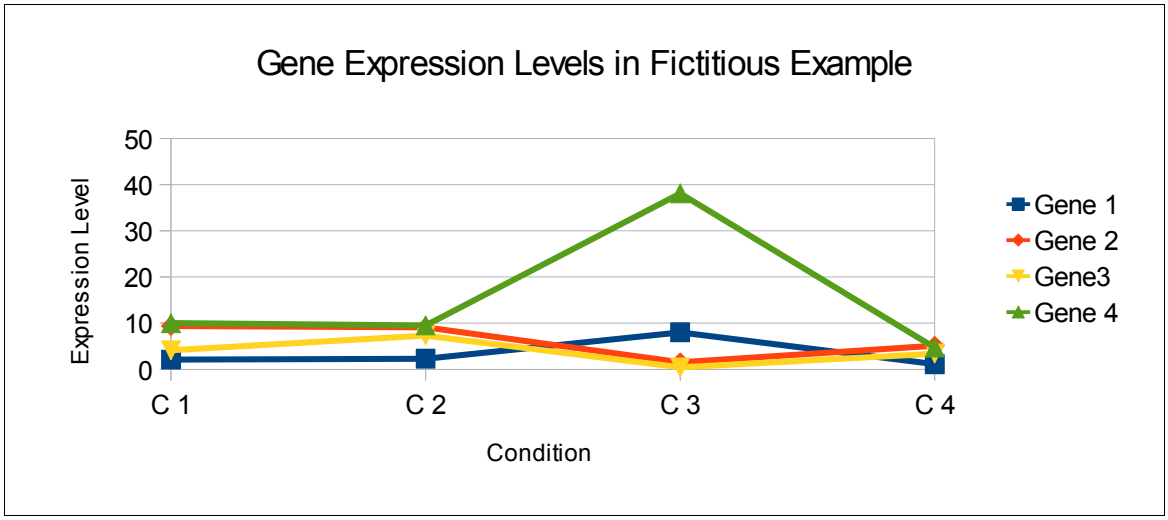

Figure 1: A simple example showing gene expression "shape".

From the graphical form of the data, we can see two basic shapes and identify C3 as the most interesting condition. Gene 1 and Gene 4 both have a shape with a spike at C3 while Gene 2 and Gene 3 have a shape with a dip at C3. Clustering in this fictitious example would allow us to group these genes into two separate groups (clusters) of expression. The first cluster of Gene 1 and Gene 4 is upregulated at C3. The second cluster of Gene 2 and Gene 3 is downregulated at C3.

The following two examples will attempt to further clarify the meaning of these clusters. In one example, the conditions are new compounds under study for possible use in a cancer drug. C1 and C2 are controls, while C3 and C4 are active compounds. This will show us that the C3 drug, which may be targeting Gene 4 only, is affecting other genes which may share some activation pathways, causing possible complications. The attempt to adjust the expression of Gene 4 may also adjust the expression of the other genes, resulting in possible side effects. (With a small and fictitious set of data, determining the significance of the variability in the other genes is difficult and would be a large concern if this were a real instead of theoretical experiment.)

In the second example, suppose that the conditions are mice that were all observed in a test looking for response to light. Mouse C3 was the most likely to ignore a bright light in order to attain food, while the other mice would not approach the bright light to get the food. The gene expression data would help to see what factors may be at play. We may know that Gene 4 is linked to a strong food response, and Gene 2 is linked to a strong fear response, but the other genes are unknown. This allows us to see these genes as possible connections in a complex system such as mouse behavior.

Analytical clustering over a large gene set (or an entire genome) can be done in many ways. The different algorithms available usually work in one of two ways: partitioning the genes into a certain number of clusters, or hierarchical clustering that creates a tree structure and defines a relationship between clusters (D'haeseleer, 2005). The algorithms we implemented into Galaxy were taken from the C Clustering Library (De Hoon, et al., 2010).

The partitioning algorithms used were k-means and k-medians. Essentially, the user defines the number of clusters, k, for either case. The difference between the algorithms lies in how they calculate the center of the cluster, via mean or median. To understand the center of a cluster and how it leads to partitioning, we must treat each gene's data as a point on a graph.

The graph in Figure 2 shows another set of fictitious gene expression data. Each point represents a gene, and each axis an experimental condition. The coordinates correspond to the

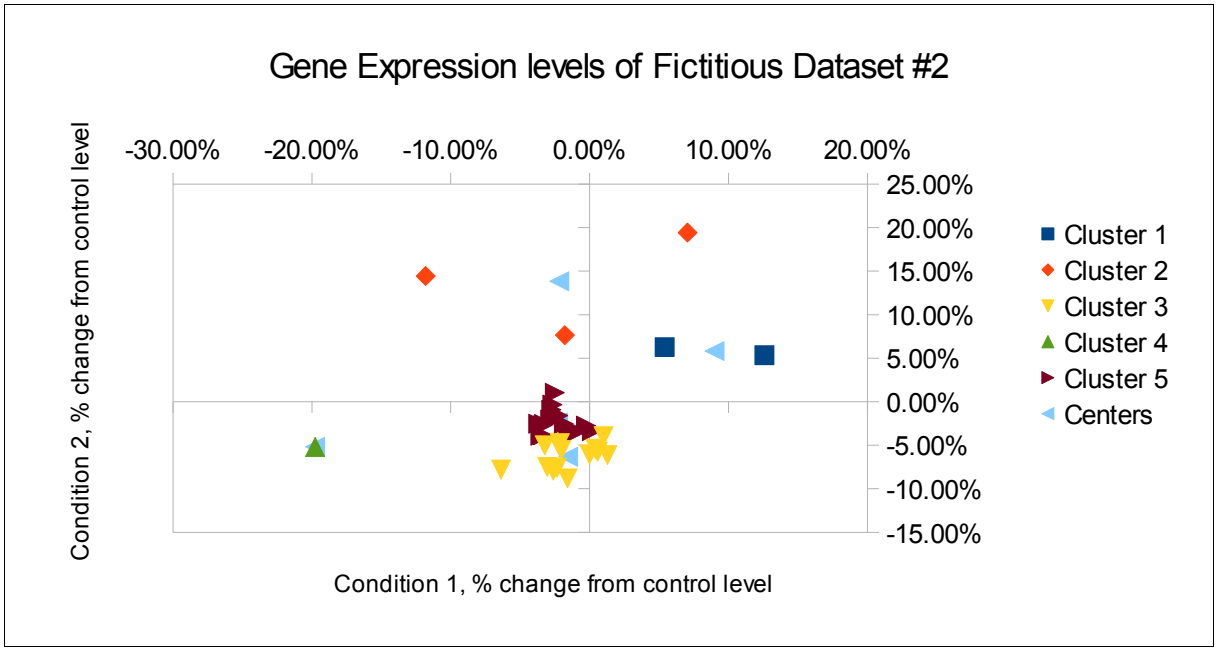

Figure 2. A graph showing the concept of clusters. There are 2 experimental conditions and therefore 2 axes. Each point represents a gene. The placement of each point corresponds to its expression level on each axis (condition). This clustering is the result of a k-means algorithm set to define 5 clusters.

expression level of that gene at that condition. On a graph interface we can conceptualize the clusters. This 2 condition graph exists in 2 dimensions, but the fundamentals can be expanded to n conditions and n dimensions.

This snippet of data has been normalized to a baseline control to show experimental upregulation or downregulation, and thus included negative values. The two condition-axes meet at the origin, which is the control level. The clusters in Figure 2 were calculated using Euclidean distance, which is evident in the graph. Essentially, cluster centers are placed to minimize the distance between points on the graph and the center of the cluster with which they are grouped, which is also the shortest distance. For example, the gene at (.054, .063) is represented by a blue square, and has a Euclidean distance of ~.036 to the cluster center at (.090, .058).

In order to achieve a similar result with the non-normalized data, we use a measure of correlation. Essentially, a correlation can be perceived as a line that approximates the points of a set when graphically plotted. A correlation of 1 is a positive-sloped line that exactly describes all points in a set, i.e. they all sit on the line. A correlation of 0 is random. A depiction of a non-normalized data set is seen in Figure 3.

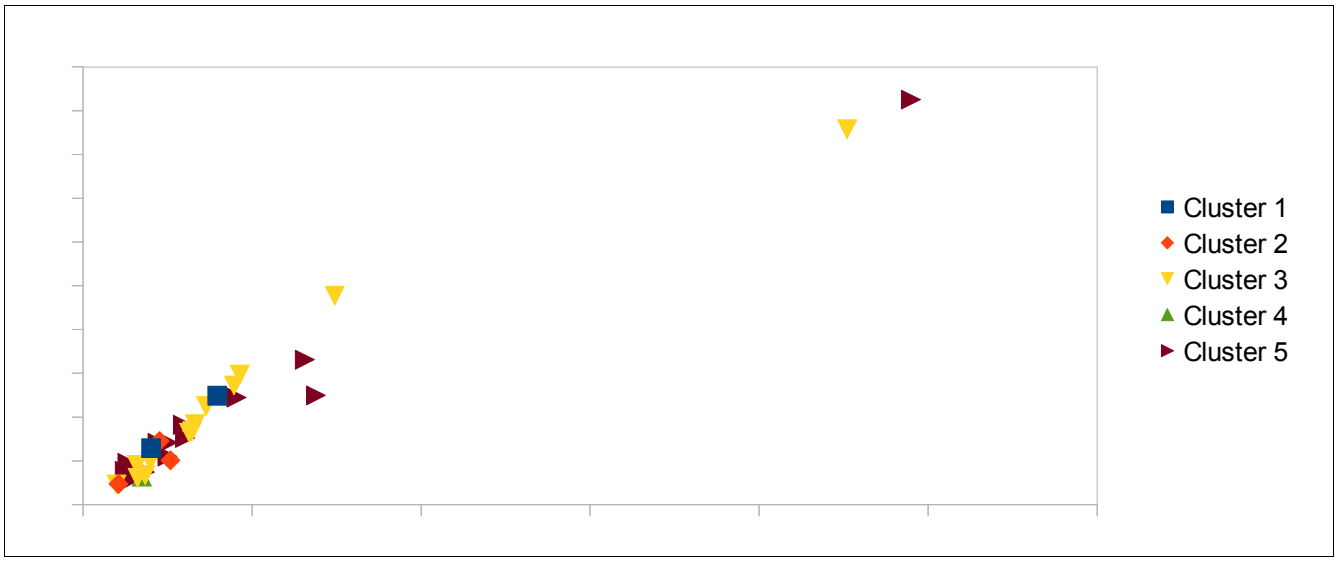

Figure 3. The clusters from Figure 2 without normalization. Clustering a set of data like this requires a measurement of correlation, such as Pearson's.

We had to use a correlation for the raw data because of the extra information included in the data set. Specifically, this set has expression measured in terms of quantity of RNA. Our clustering is done only to compare the delta, or change in expression levels; the quantity is irrelevant.

Comparing Figure 3 to Figure 2 can illustrate the difference. In Figure 2, data is plotted in terms of change. Every gene that is upregulated in both conditions would be in the positive-x, positive-y quadrant. The ones that are more upregulated are further out from the origin and closer to the other genes with similarly large levels of upregulation. The Euclidean metric groups these genes together in space on the graph to cluster them. Applying that type of algorithm to the graph created in Figure 3 would produce a vastly different structure. Since all points on the graph are actual RNA levels under each condition, no negative values exist. The points that are furthest from the origin do not share a

similar high level of upregulation, but rather a high level of RNA. A Euclidean clustering would pair the two genes furthest away from the origin, which is not necessarily correct in this case.

While the Euclidean metric operates on a spatial level (clusters are gathered together in graph-space), correlations gather points into clusters on a linear level. Suppose the following: Five genes all have expression levels twice as high in Condition y than in Condition x. Regardless of the base level of RNA produced under x, there will be twice as much of it under y. A line could be drawn through the graph at y=2x that would correctly cluster these five similar genes together. Instead of gathering points together in space, correlations draw lines that approximate the data (Figure 4).

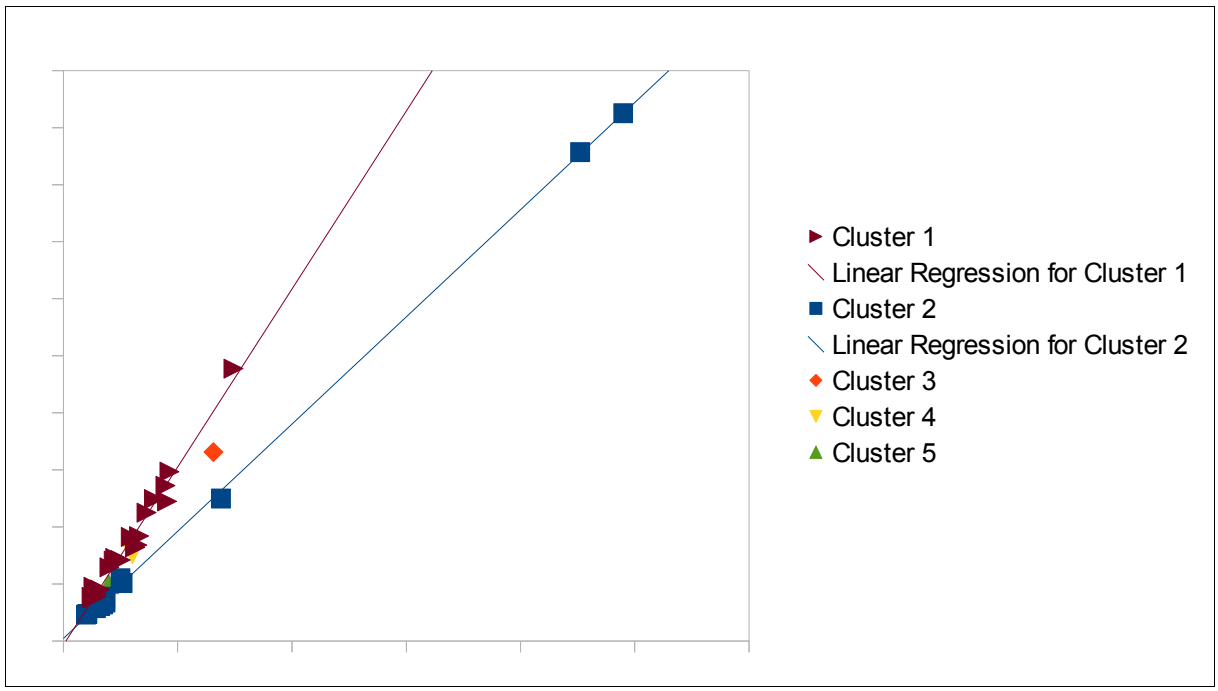

Figure 4. Data from Figure 3 re-clustered to show correlation calculation. Clusters 3, 4, and 5 are single points. The linear regression lines for clusters 1 and 2 demonstrate how a line can approximate a cluster.

The correlation algorithms provided by the C Clustering Library are the Pearson correlation coefficient, Absolute Pearson correlation, uncentered correlation, absolute uncentered correlation, Spearman rank correlation, and Kendall's τ. Also provided are the metrics Euclidean distance and Manhattan (or city-block) distance. See Appendix A for algorithms.

For a hierarchical clustering, we need to define measurements for not only the distance between gene expression points, but the distance between clusters. For measuring this distance, there are multiple options; the C Clustering Library provided four. A single-linkage measurement is the shortest possible distance between any of the nodes in the first cluster paired with any of the nodes in the second cluster. Maximum linkage (also called complete-linkage) is the opposite, taking the longest possible distance. Average-linkage is simply the average of all pairing combinations between the nodes of the two clusters. Finally, centroid-linkage measures from cluster centers like the ones shown in Figure 2.

Clusters define a similarity between the gene expression data and the hierarchical tree defines the similarities between the clusters. The algorithm starts by creating a node connecting the two closest clusters. This node is also treated as its own cluster. Recursion then matches the next closest clusters, and so on until all clusters are connected. Once there is a tree, in order to visualize it to see the patterns

apparent in the data, a tool that is commonly used is the heat map (Figure 5). These colored matrices are typically three colors, green for down-regulation, red for up-regulation, and black for no change. Each row in the matrix represents a gene. Each column is a condition or experiment. The shade of the pixel at the intersection of the two is determined by the expression value of that gene in that condition, mixing the black with the directional color accordingly.

The C Clustering Library provides a format compatible with Treeview, a Java program that visualizes statistical data into heat maps among other formats (Saldanha, 2004). Being built in Java, Treeview can be run as an applet inside a browser. Galaxy also makes its interface available via a web browser, using XML on the back-end to access its modular toolbox. Our program combines the three to make a single, user-friendly tool.

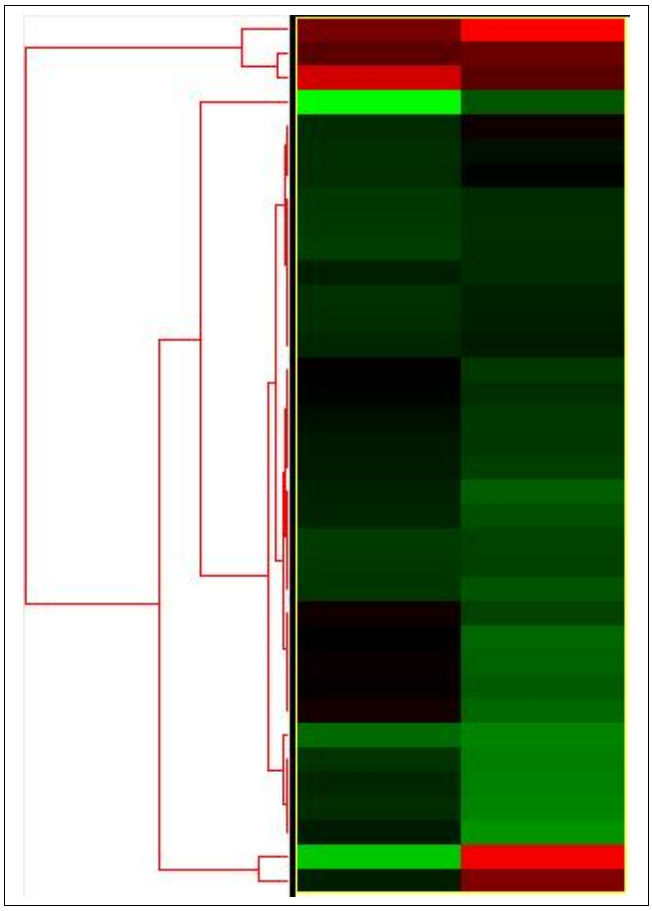

Figure 5. A heat map of the data in Figure 2. The tree on the left side represents the hierarchical clustering.

HDAC1 and its Role in Neuropathy

In order to test the new program, we worked with Dr. Mira Jakovcevski at Umass Medical School. She provided us with microarray data to be analyzed. The experiment at hand involved mice, learning and memory, and histone acetylation.

Chromosomes carry DNA, which is comprised of a double-helix of nucleotides. That double-helix wraps around histone proteins to form a building block called the nucleosome, described as being similar to a spool of thread (Backstage... 2003) (Figure 6). These nucleosomes are the basic pieces of chromatin, the structure of the chromosome.

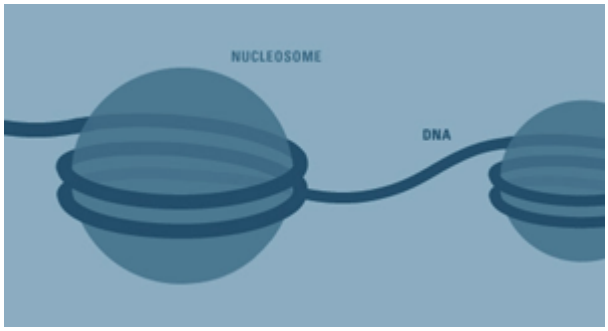

Figure 6. Depiction of nucleosomes wrapped with DNA. Picture courtesy Rockefeller University.

Histone proteins play a role in gene expression, and acetylation of those proteins has an effect about which we are still learning (Fischer, et al., 2010).The acetylation of histone proteins is the attaching of an acetyl group (COCH3) to an exposed tail. This process is controlled by the action of histone acetyltransferases (HATs). The reverse process removes the acetyl group from the histone by the action of a histone deacetylase (HDAC) (Fischer, et al., 2010). A tail with an acetyl group is associated with a gene undergoing transcription. Histones with

bare, unmodified tails are found in a closed form of chromatin that does not have DNA available for transcription (Holliday, 2006) (Figure 7).

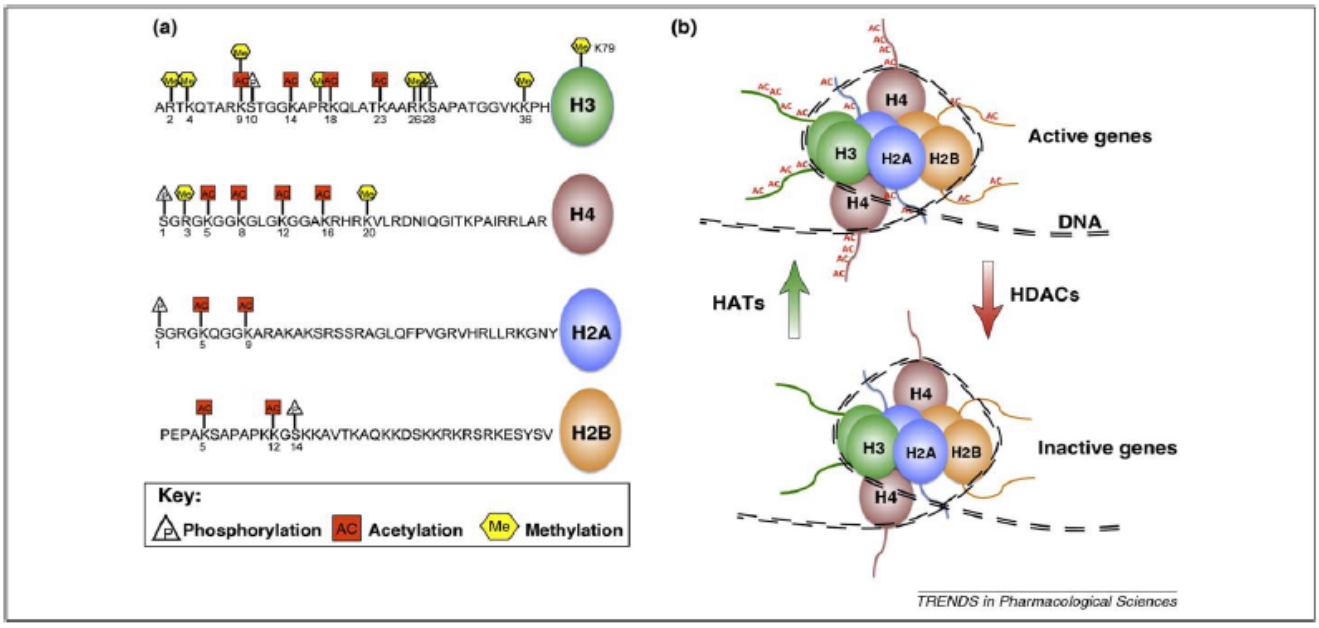

Figure 1. Histone modifications regulate gene expression
(a) The histone tails are subjected to massive post-translational modification, including phosphorylation, acetylation, and methylation. Those modifications give rise to a distinct pattern, also called the 'histone-code', that affects gene expression. (b) Acetylation of lysine residues on histone tails is usually associated with active genes, whereas reduced or no acetylation is found at inactive genes. Histone acetylation is regulated by the counteracting activity of histone acetyltransferases (HATs) and histone deacetylases (HDACs).

***Figure 7. Effects of acetylation/deacetylation on gene activation. A nucleosome is comprised of the histone proteins H2A, H2B, H3 and H4, shown in (a). A nucleosome containing histones in both states is shown in (b). Courtesy* Trends in Pharmacological Sciences.**

Increasing histone acetylation has been shown to have an effect on neurological function, significantly increasing cognition and memory (Kazantsev and Thomson, 2008). As acetylation is a sign of healthy neurological function, deacetylation (and therefore HDAC action) is often an actor in neuropathy. HDAC inhibitors have been shown to repair cognitive function in cases where HDAC regulation was a factor in the pathology (Guan J, et al., 2009). In that study, they showed that the upregulation of HDAC2 played a significant role in inhibiting spatial learning. They did not find the same with upregulated HDAC1.

The overall goal we seek is to increase the amount of information available about the HDAC family, since they have shown promise in possible treatments for some neurological disorders. Broad-based HDAC inhibitors are worrisome without knowing more about the discrete functions of the HDAC enzymes. The best overall outcome is the development of HDAC inhibitor treatments targeted specifically to the enzyme actor involved in the neuropathy.

Methods

Clustering Tool Development

Integration of a clustering and visualization tool into the Galaxy framework relied upon a stable development environment. The testbed to be used was a personal laptop (x86 Toshiba PC, A305-S6858) running Windows Vista Home Premium. Galaxy runs off a combination of C and Python code, so initial prep of the laptop included installation of Python and Microsoft Visual Studio Express for its Visual C library and compiler.

Windows was chosen as the original development and test environment out of a desire for simplicity and useability. It proved difficult to configure even with the tweaks recommended in the Galaxy documentation. Linux was chosen as the next development environment for its native C and Python capabilities.

The laptop was prepared by installation of Oracle's VM VirtualBox. This allows for a computer to run two concurrent operating systems. A Linux sub-machine (or virtual machine) was established using a Gentoo distribution. The kernel build was manual and minimal in order to minimize time and resource use.

Galaxy is updated and synced via the Mercurial revision control tool. This allows for updates written for the software to be easily applied to a local server. Mercurial was installed onto the virtual machine and with it, Galaxy-Central was installed. There are two versions of Galaxy. Galaxy-Central is the developer version. Galaxy-Dist is the end-user version.

With a working Galaxy instance available on the virtual machine, the next step was to enable communication between the Linux and Windows systems, so that a user could interact and run programs using the standard Windows interface. Galaxy communicates through an internet interface, launching tools/programs on the webserver with xml definition files. From the user standpoint, a web browser pointed to the correct address is the depth of complexity involved in access. Galaxy's built in webserver by default serves its homepage to 127.0.0.1:8080. Even though the same machine contained the webserver and the client browser, the browser wouldn't recognize the address as typed. This problem was solved with the virtual networking capabilities of VirtualBox, causing 127.0.0.1 on the host windows machine to be pointed to the local service of the Gentoo virtual machine. With this forwarding in place, the configuration of the testbed was complete.

Building a tool in Galaxy required 3 steps: the creation of a python program file, the creation of a tool definition xml file, and the editing of the tool_conf.xml file to register the new tool with Galaxy. We first created a skeleton placeholder xml file and registered it with Galaxy to enable testing with the interface.

The python program file needed to be able to take an input file, cluster the results, and return to Galaxy with a visualization of the data. The math to cluster the results was not the goal of this project, and we decided to use the existing algorithms contained in the C Clustering Library (De Hoon, et al. 2010). The Pycluster module was downloaded and copied into the Python library on the virtual machine. With this in place, the algorithms were available to us with a simple call of:

```
from Pycluster import *
```

The clustering we were most interested in was hierarchical. The algorithm available for hierarchical clustering had parameters available that would define how the calculations were to be done. It allowed for selection of linkage type (single, maximum, average, or centroid) and distance evaluation (Pearson's correlation, absolute Pearson correlation, uncentered correlation, absolute uncentered correlation, Spearman rank correlation, Kendall's τ, Euclidean distance, and city-block distance, see Appendix A for more information.) We decided to implement all possible options to leave the choice to the user.

In order to give the user an option, we used the sys.argv list to read parameters sent into the program. The xml tool definition file was edited to define the command line of:

```
prep4jtv.py $infile $outfile $method $dist $xcluster $xmethod $xdist
```

Prep4jtv.py is the name of the python program file, method allows selection of the linkage type, dist allows selection of the distance evaluation, and xcluster is true if the conditions are being clustered as well as the genes, with selectable options for that clustering as well.

The output of the algorithms are lists or arrays of numbers, and there was no built-in visualization. The documentation for the C Clustering Library refers to Java based Treeview as a separate program that does tree visualization. The library has an object (called Record) that writes the algorithm output into files of the proper format for Treeview.

The Record object became the basis of the program. It reads in the data from $infile, creates a hierarchical clustering of genes according to the input parameters, creates a hierarchical clustering of conditions accordingly, and makes output files. Treeview takes in a base file of filename.cdt (cluster data). This file has the position of each gene and experimental attribute, as well as the value of the genetic expression at that attribute. If available, it also takes in filename.gtr (gene tree) and filename.atr (attribute tree.) These files define the relationship between the rows and between the columns, respectively. Treeview looks for these two optional files in the same file directory as the .cdt file. If a matching filename.gtr or filename.atr are not in that directory, the program assumes they don't exist. When Galaxy saves outputs, each output file is stored with an internal reference, in its own directory. The file output from a tool may be “output.gtr” but Galaxy will save it as dataset_#somenumber#.dat.

Treeview as a standalone program does not fulfill the requirements of something able to be launched from any web browser. It has interactivity that allows a user to search for a gene or select a portion of the available cluster for a closer look. This level of interactivity would make run-and-return scripts in the browser slow and not very user friendly. Being based in Java, however, allows the program to be installed only on the webserver that hosts Galaxy and run as an applet in most web browsers. This acts as a full program in regards to interactivity, but applets have some limitations for security purposes. Applets cannot access files directly to read or save. The only option is for the web page that launches the applet to pass a file with it. Treeview recognizes a tag for this purpose:

```
<param name="cdtFile" value="...">
```

The ability to pass in the other two files with the tree data does not exist if they are in separate directories, as Galaxy saves them. To ensure the proper placement of the other files, we used os.system calls after the files were saved to move and rename the files to the same directory with the

same base filename and proper extensions. This ensured that the clustering algorithm was saving the files in the proper format to be read by the Treeview tool.

Installing Treeview as an applet was possible in the galaxy-central/static folder, which ensures a file can be accessed with a reference to its file location. Galaxy does not have a native datatype that launches applets, but it does support an html file that has access to the galaxy-central/static folder. Our python file was edited to create another file along with the cluster data files. It writes a basic html file with an embedded applet, passing in the proper parameters describing the correct location of the saved cluster data files. This allows for one tool to take the input file, do the calculations, save the output file, and then pass that output file to a visualization applet that offers interactivity with the data. See Appendix B for source code.

Gene Expression Data Analysis

The gene expression data from the mouse microarray experiment spanned 35,518 genes, each on its own row, with the eight individual mice on the columns as the experimental conditions. While this amount of data could be clustered by a computer, the computational time would be heavy and likely not worthwhile. We created a mean-set for each group (experimental and control) with the average expression values and standard deviation at each gene. We then used a Student's t-test to determine differential expression of the means. A standard $p < 0.05$ produced only a handful of genes, far too few to reliably cluster. We increased the p value to $p < 0.2$. Of the 36k set, 235 genes showed that level of differential expression, leaving us with a much more manageable set of data. See Appendix C for this set of data.

The clustering and visualization tools were run four times for this dataset. By using different clustering algorithms or normalizations, we had more chances to see something significant and to verify earlier findings. Each time, we used average-linkage for the method of calculating cluster distances. Average-linkage checks the distances between all pairings of nodes in a cluster and averages them to find the cluster distance. This allows for a calculation less affected by a possible outlier.

For the first clustering run, the data were adjusted for scale by converting the expression levels to a $\log_2$ scale. A baseline was created by taking the data from the 4 control mice (MJ5, 7, 9 and 11) and calculating the mean for each row/gene. This gene expression mean was then subtracted from each gene. This left the genes that were upregulated from the baseline with a positive value, which would show up red on the heat map. The negative genes are downregulated, and show up green on the heat map. This data was run through the clustering tool with the Euclidean distance parameter.

In preparation for the second run, the data were normalized on a per-row basis to a zero-sum percentage change from the mean. (Each cell was divided by the mean for the row. When normalized like this, the sum of the upregulations in each row equals the sum of the downregulations.) These were also run with the Euclidean distance parameter.

For the third run, we used the raw data without normalization. We chose the Pearson correlation algorithm for this set of data because it handles variations in average expression level well (D'haeseleer P, 2005).

In order to verify negative correlations in our findings, we ran a fourth test with the raw data. This run was an absolute Pearson correlation, which allows for the clustering to disregard the sign of the correlation number, intermingling positive correlations with inverse correlations.

Results

In order to provide Galaxy users with a tool for gene expression clustering and visualization, we chose the algorithms provided in the C Clustering Library and implemented a tool in Python. This tool creates an HTML document with an embedded Applet that launches the Java program Treeview. The Treeview program allows interactive visual browsing of the data.

We were provided with microarray data corresponding to eight mice: four with upregulated *LacZ* (control) and four with upregulated *HDAC1* (experimental). The upregulation was localized to the cingulate cortex, which was later dissected to provide the material used in the microarrays. These data were analyzed with the new clustering tool and visualized to identify gene expression patterns and their correlation with *HDAC1*, either positive or negative. The resulting genes are to be targets for further study in pursuit of greater understanding of neurological disorders and the amelioration of HDAC inhibitor side effects.

New Clustering Tool

A script was written in Python that performs clustering and prepares the data for visualization in a Treeview applet. We named the program prep4jtv.py. The source code for the program and the xml file that incorporates it into Galaxy can be seen in Appendix B.

Launching the program begins with an open instance of Galaxy. Figure 8 shows the initial state of Galaxy. The user first needs to give Galaxy the data on which the tool will run. Our data was uploaded from the local computer. This is done by clicking the Get Data tool link, and then Upload File. A user can then select the file off the local drive with the Browse button (Figure 9). Once a file has been successfully uploaded it will appear in the right History panel (Figure 10).

The file format that our clustering tool will recognize is a basic tab- or space-delimited table in which the first row is attribute labels and the first column is gene labels. There can be no blank cells, not even the cell preceding the attribute labels, and each piece of data must be numerical (no scientific notation.)

Once the data is available to Galaxy, the clustering tool can be launched. The tool family is Statistics, and the link is Java Treeview (Figure 11). (In a future revision, this link will move to the Graph / Display Data tool heading.) The user now selects the file from the History pane to cluster (by default, the most recent file is selected). If the columns or experimental conditions are to be clustered as well as the genes, the user checks the checkbox. In experiments with control sets and experimental sets, they should cluster together and highlight outliers. In the case of the data we used, each column was a mouse. We expected to see them clustered into two parts according to upregulated *LacZ* (control) or upregulated *HDAC1* (experimental).

The other selection boxes allow the user to select the Clustering Method (linkage) and the Distance Function. There is a selection box for the gene set and the condition set, to give the user the option of calculating columns differently than the rows. The Linkage methods implemented are: single-linkage, maximum-linkage (or complete-linkage), centroid-linkage, and average-linkage. The Distance Functions implemented are: (Pearson's) correlation, absolute correlation, uncentered correlation, absolute uncentered correlation, Spearman's correlation, Kendall's τ, Euclidean distance, and city-block distance. (Linkage methods are discussed in the Background section, and distance functions are defined in Appendix A.)

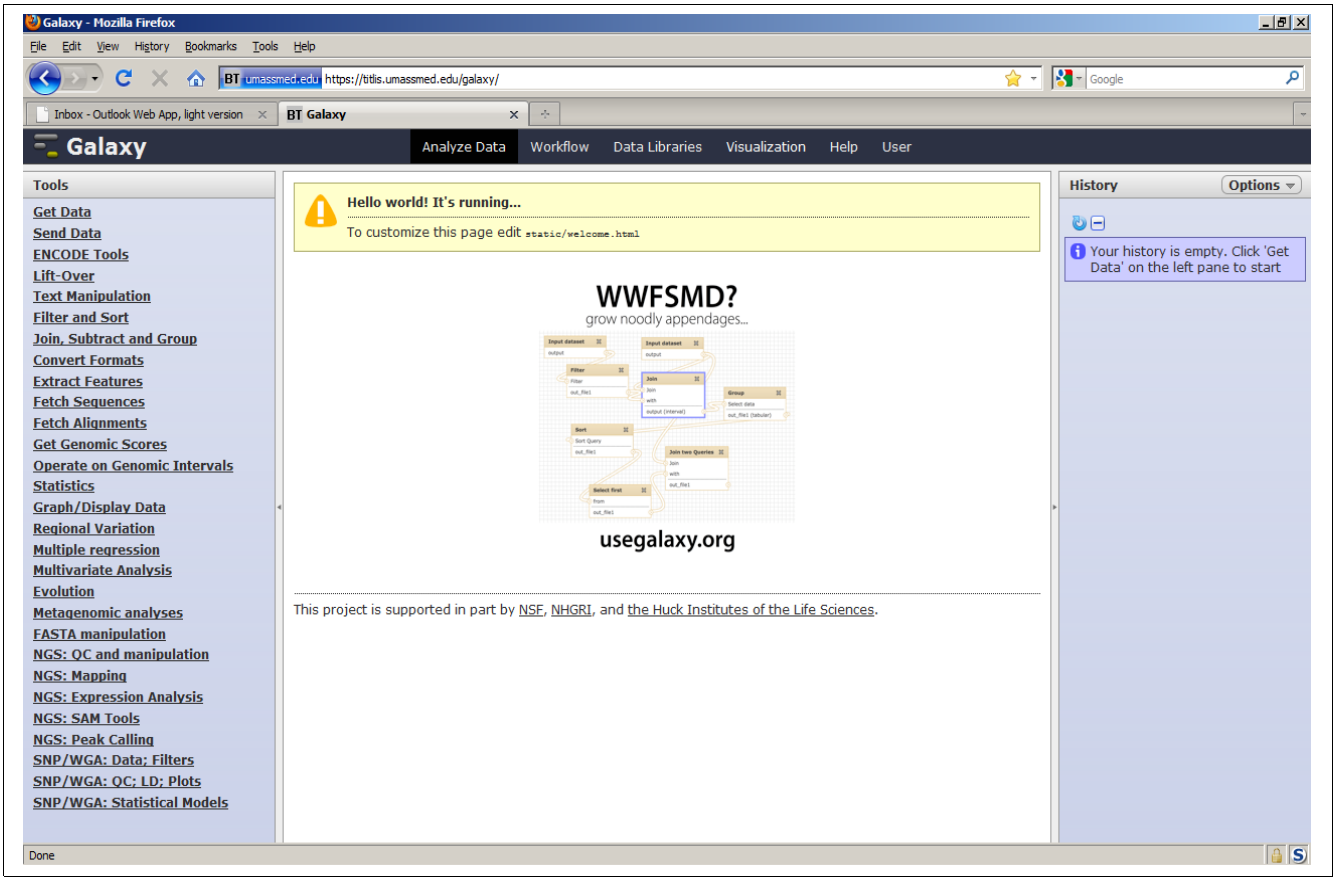

Figure 8. The Galaxy interface. The blue panel on the left shows all the families of tools. Clicking one of the links will open up the tools underneath that heading. The middle panel is where the tool interface will launch once selected. The right blue panel contains the data as it is manipulated. This history is empty as it contains no data yet.

Once the arguments for the algorithm are selected and the execute button is pushed, the script runs and will return a file to the History pane (Figure 12). This file in our implementation is webpage (HTML file) with an embedded applet. In order to access that page to launch the applet, the user clicks on the "eye" icon of the "Clustering..." file. It posts to the center frame with a button labeled "View results.cdt" (Figure 13). Clicking that button launches Treeview in an applet. This provides a powerful visualization tool to the user without any additional software needing to be installed (provided the browser is Java capable).

Treeview is a program that is freely available on the internet. We did not author or alter the program, but simply introduced it into Galaxy as an applet. Using Treeview was the best choice for the limited development time available, and has the added benefit that some users are likely to be already familiar with the basic functions of the program. Not all functions of the full program are available because of the restrictions on applets. Documentation is available in Appendix F.

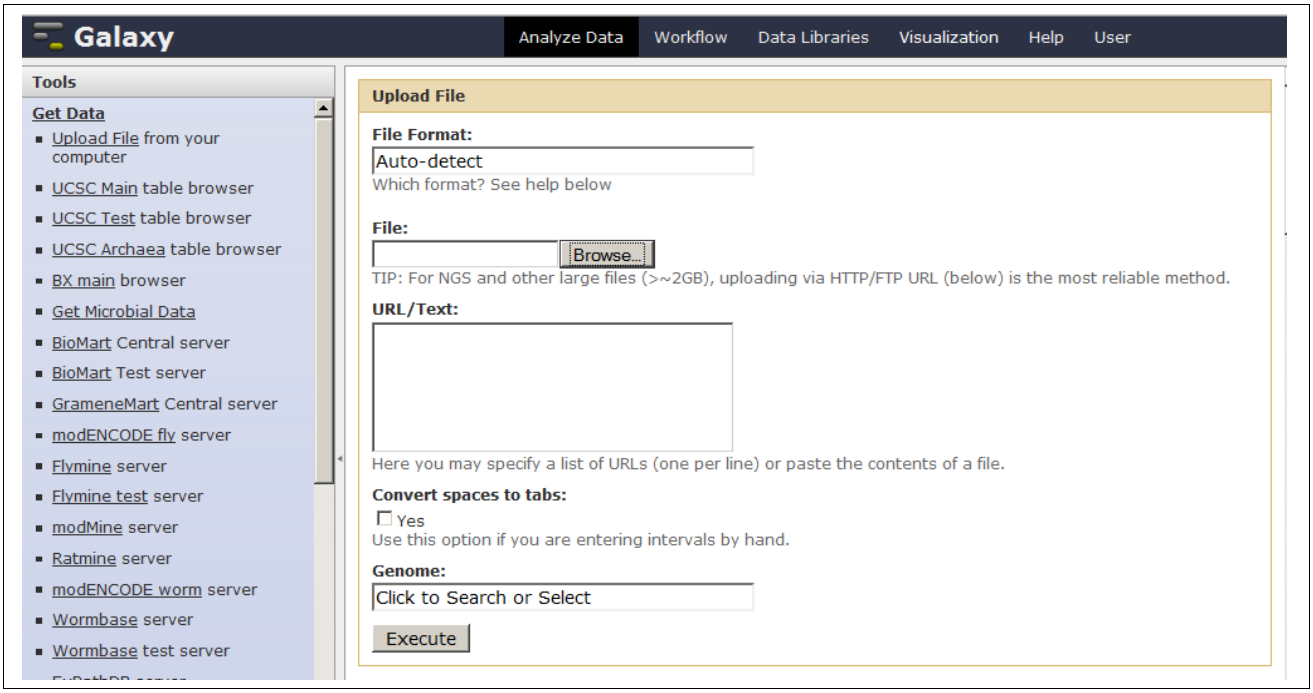

Figure 9. Galaxy Upload File interface. The left panel shows the tool navigation: we've selected Get Data as the tool family, which then lists all the ways to get data. The middle panel shows the Upload File tool interface. The browse button is highlighted. This explores the local computer to select the file.

Figure 10. Galaxy History pane showing an uploaded file. Clicking the name will expand the item with file details and a preview. The eye icon will display the file in the middle Galaxy pane, the pencil will allow for attribute edits, and the X deletes the file.

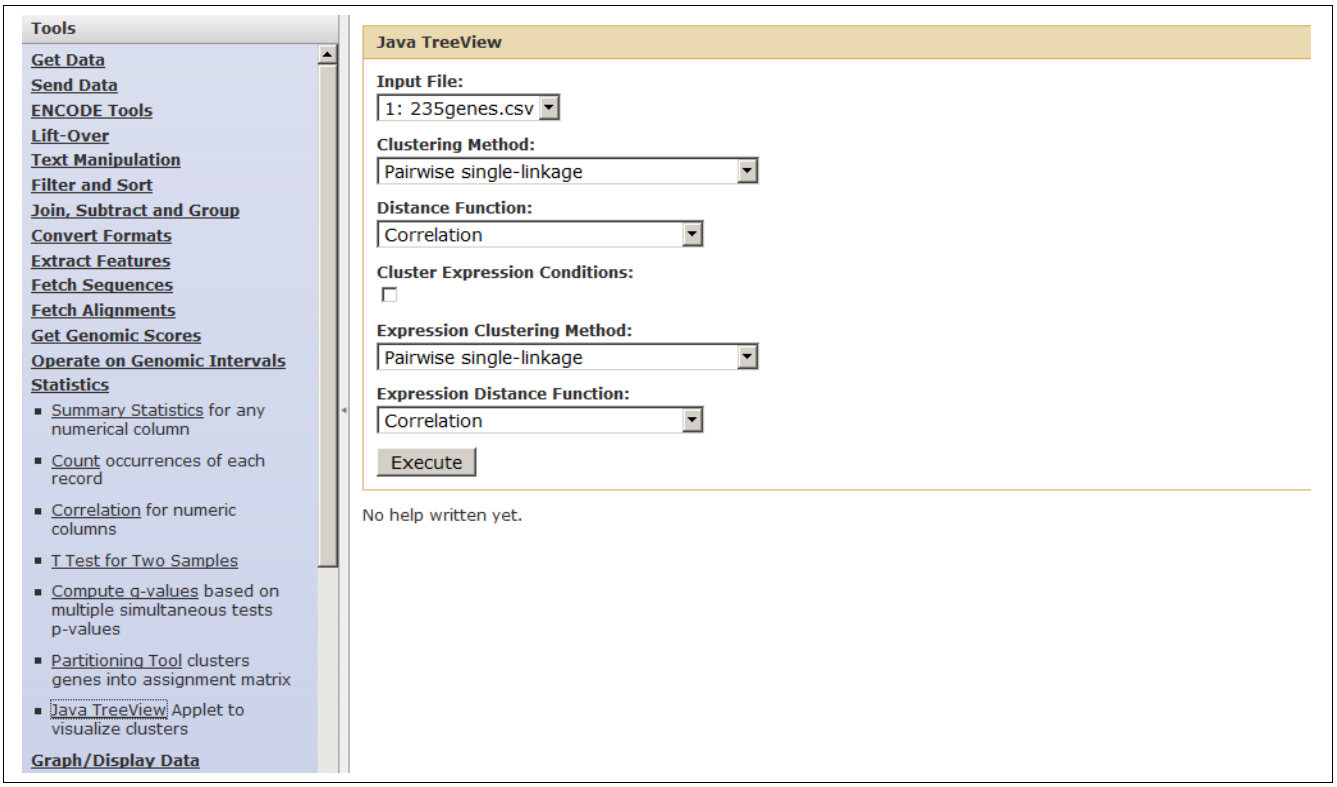

Figure 11. Clustering tool within Galaxy. The middle pane shows our tool interface. The Input File is a dropdown selection box that imports datasets from the History pane. The Clustering Methods allow for cluster linkage selection. The Distance Functions allow the user the choice between 6 corelations and 2 metrics. The checkbox for Cluster Expression Conditions allows the user to create a hierarchical tree for the attributes (experimental conditions). If left unchecked, the tool will only create a tree relating the genes to each other.

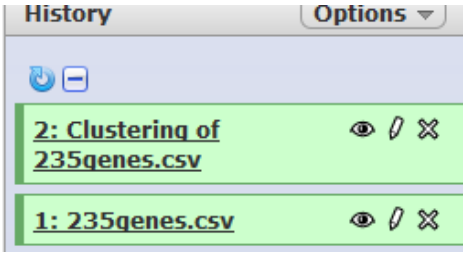

Figure 12. History pane showing the newly created file, output by the clustering program.

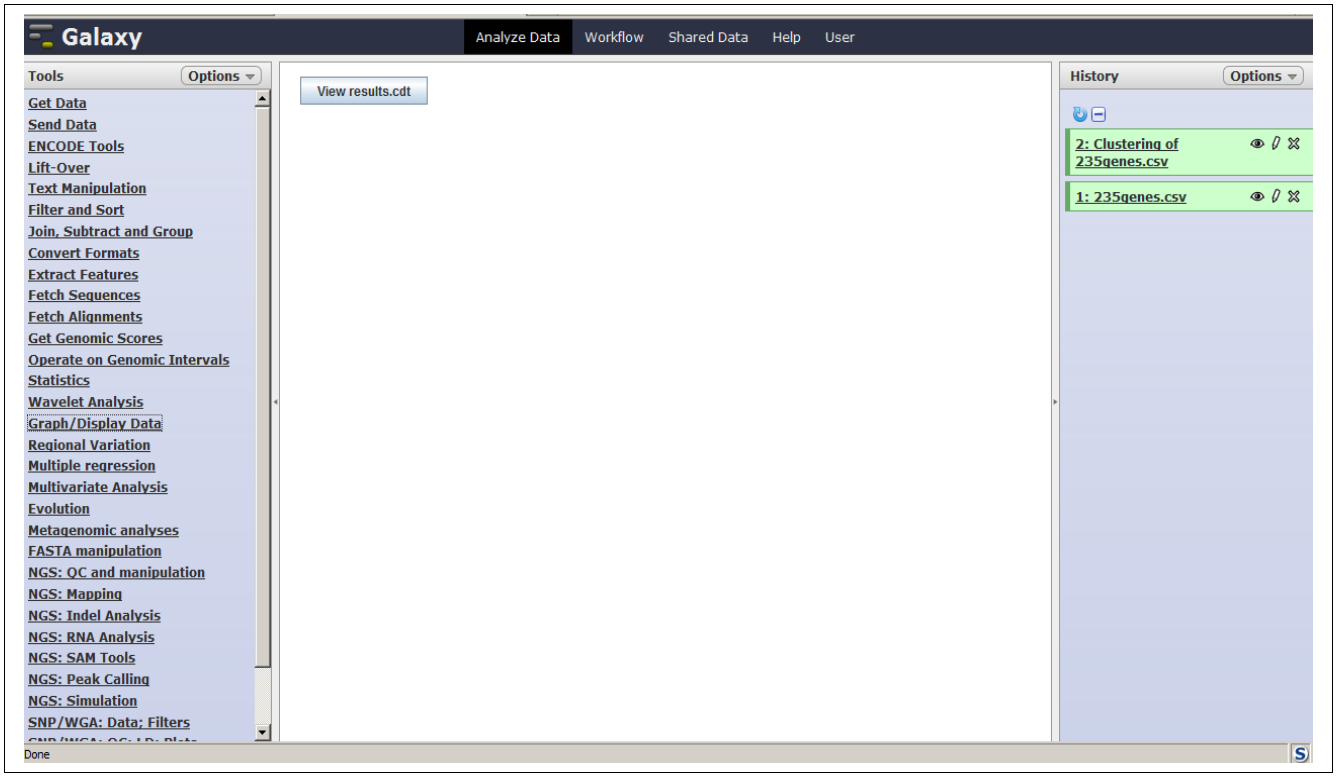

Figure 13. Galaxy displaying the output page from running the new clustering tool. After clicking the "eye" icon on the "Clustering..." file, the middle pane shows this page with a single button. Clicking this button launches Treeview.

Analysis of Microarray Data

The provided microarray data were run through the clustering and visualization tools three times. The first time, with a binary log normalization, can be seen in Figure 14. This data was normalized to define normal *HDAC1* expression as the mean of the control values. The first test of the cluster was the attribute cluster across the top of the middle panel. The control mice are labeled MJ5, 7, 9, and 11. The experimental mice are MJ10, 12, 14, and 22. As expected, the mice were correctly partitioned into the experimental *HDAC1* set and the control set. Identifying *HDAC1* in the gene set was done by referencing the labels in Appendix C. The identifier for our target gene was 10401136. This number can be used as a search term within Treeview, under the Analysis drop-down menu. It would have to be formatted as 1.0401136 however, as seen in the frame on the right of the program.

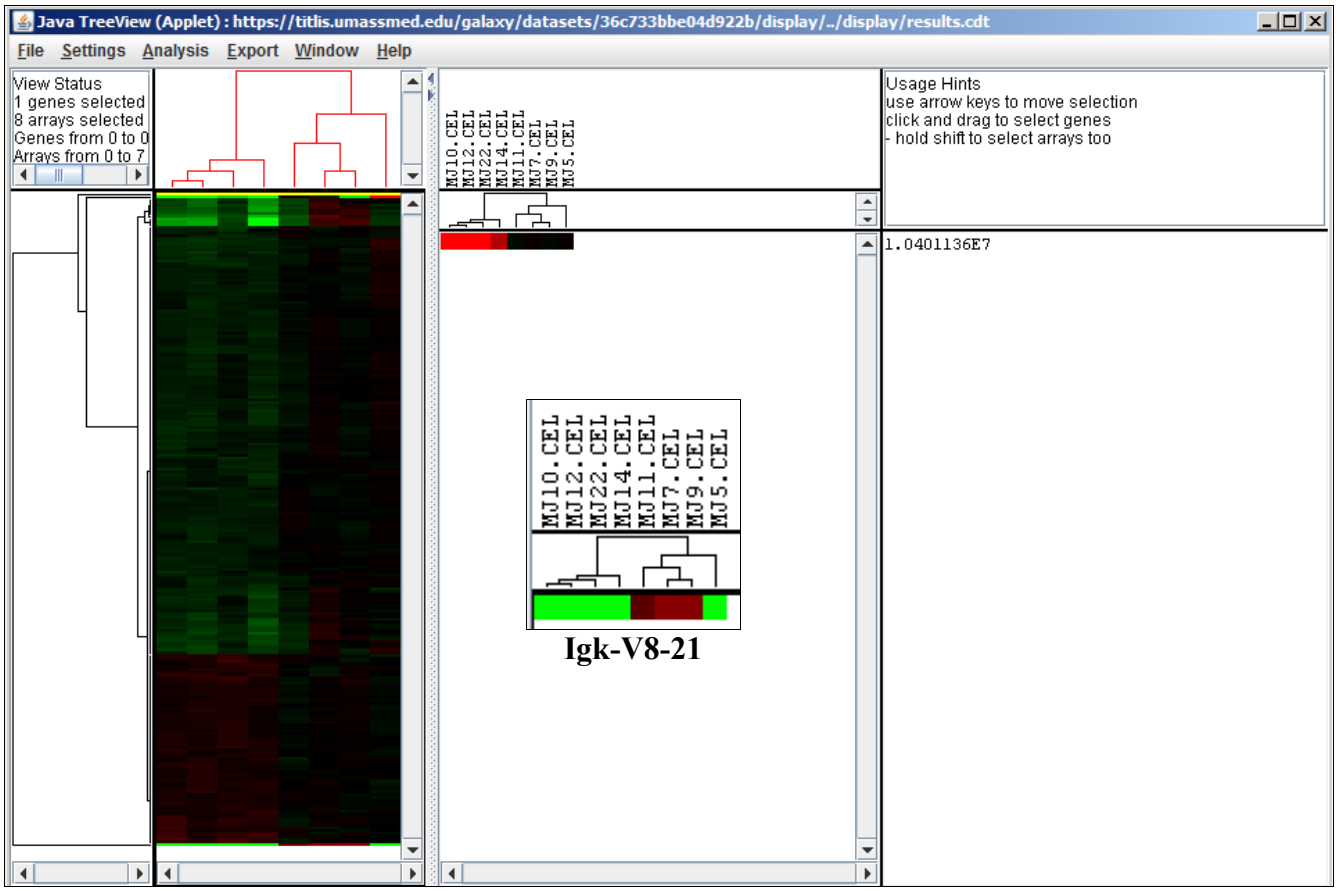

Figure 14. Treeview visualization of Log-2 normalized microarray data. The far left pane contains the hierarchical gene tree. The left pane contains the heat map with the expression data. The middle pane shows an enlarged selection of the heat map, with the condition cluster across the top. The top row in the heat map, corresponding to* HDAC1, *is selected in this view. As expected, the first 4 pixels, corresponding to the experimental group, show upregulation, while the second 4 pixels, corresponding to the control group, show normal regulation. The inset shows the bottom line of the heat map.

In this cluster tree, *HDAC1* was all the way to one end, at the top of our graph. Looking at the

expression data, its closest neighbor did not appear to be positively correlated, leaving *HDAC1* as an outlier in its own cluster. The dendrogram showed only one other gene further removed from the tree than *HDAC1*. At the other end (Figure 14, inset) was gene 10545239, which corresponds to the label "ENSMUST00000103387 /// M28833 /// AJ222611." ENS refers to ensembl code, MUS for mouse, and T for transcript. The number 00000103387 maps to the gene *Igk-V8-21* according to the Ensembl Genome Browser. Placing these two gene visualizations side-by-side showed a possible negative or inverse correlation, as can be seen in Figure 15.

The second run of the tools, this time on the zero-sum normalization, can be seen in Figure 16. Normalizing the data like this was an alternate to the first normalization. This run used the whole group of mice in creating a mean, instead of only the control group. This showed a clustering of the experimental mice, but the control mice did not cluster together neatly as we saw in the first run. Looking at the gene tree, we saw *HDAC1* as the third most distant gene from the cluster. The second furthest was again 10545239, or *Igk-V8-21*. The absolute furthest gene was 10454192, labeled *Ttr*.

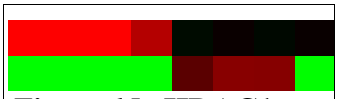

***Figure 15.* HDAC1 *(top) shown with* Igk-V8-21. *Left 4 pixels represent experimental mice, right 4 are control. With every upregulation of* HDAC1, Igk-V8-21 *is downregulated. The reverse correlation is not quite as good, as one of the control mice also has a downregulation of* Igk-V8-21.

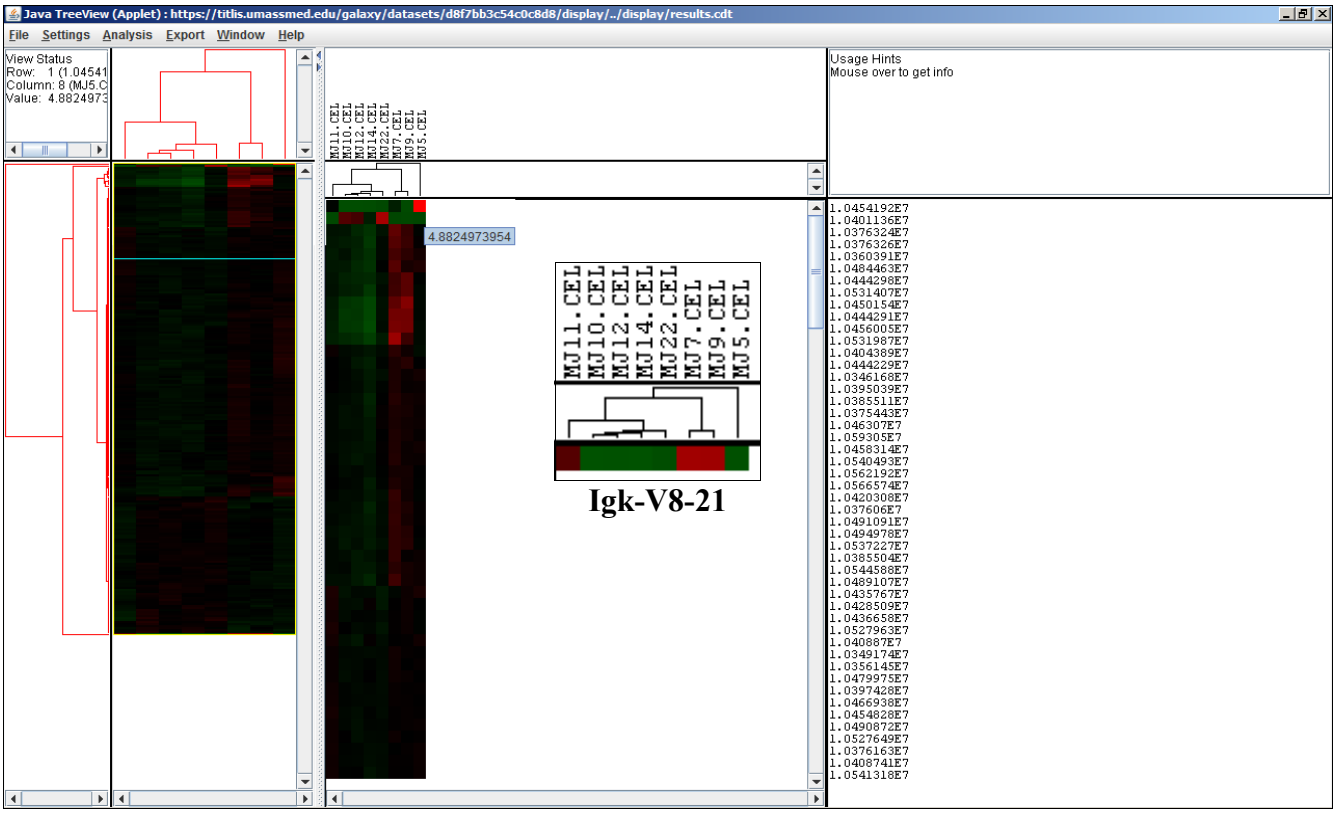

Figure 16. Treeview visualization of zero-sum normalized data. The center frame shows the top of the graph, with* Ttr *followed by* HDAC1. *The value shown is the expression level for the one mouse overexpressing* Ttr. *MJ5 had an expression level of 5.88 times the mean of* Ttr. *The inset shows the bottom of the graph.

Upon looking at the heat map, two things were immediately evident. First, only 3 of the mice showed upregulation of *HDAC1*. The expression levels had such a large variance that one of the mice in the experimental group showed a level below the mean. This normalization potentially mischaracterized the data. Second, the gene *Ttr* only showed high in one of the mice, in one of the control group. This was likely an anomaly and not statistically significant.

Rechecking Ttr in the raw data showed that in fact, the HDAC set of mice had the lowest levels of Ttr while the LacZ mice all had higher expression, some much higher (Table 2). This finding was not apparent in the heat map, but the tree clustering was able to highlight what is likely a negative regulation between HDAC1 and Ttr.

Table 2	MJ10	MJ11	MJ12	MJ14	MJ22	MJ5	MJ7	MJ9
10454192	163.85	1888.13	174.71	196.89	279.48	11677.7	1260.79	239.73

The third run of the data was not normalized (Figure 17). In this analysis we sought correlation based on change in expression levels solely; the absolute levels were not important. The heat map was all red because the data was all above a gene expression level of 3. This was expected because the data were not normalized to fit the -3 to 3 scale that Treeview uses to color its graphs. The part of the program that displays the heat map expects the numerical data to represent a relative expression level. This set of data was using the absolute expression levels, meaning the quantity of RNA on the microarray, so the heat map graphic is not useful for this type of analysis. The clustering was still accurate when done by correlation, and the trees provided a useful tool for analysis.

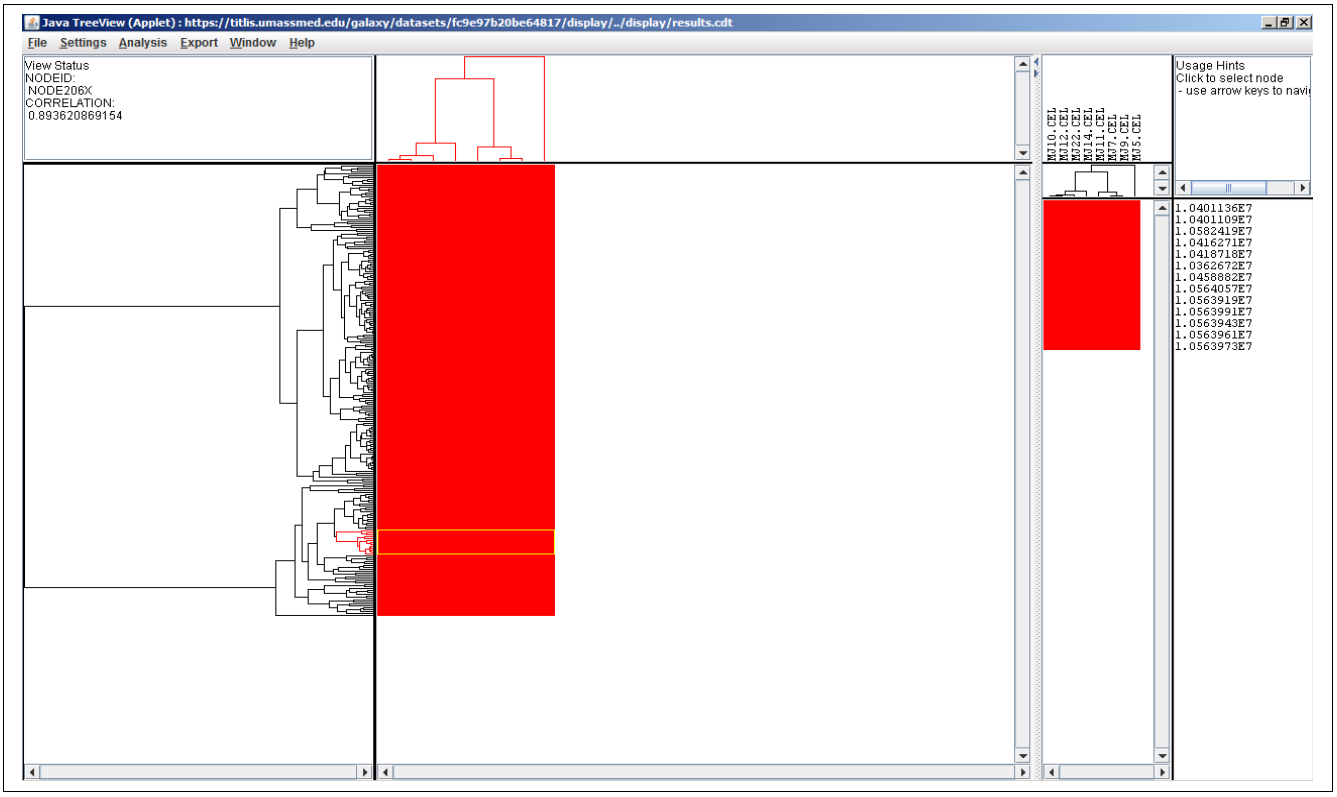

Figure 17. Treeview visualization of raw data. The frames are stretched to show the detail in the gene tree. The top-left frame shows the information for the highlighted node. The top line in the highlighted section is* HDAC1. *For greater detail see Figure 18.

In this third analysis, once again, the experimental mice were clustered together on the attribute tree across the top of the graph. Control MJ5 was an outlier, but clustered against the other controls when integrated into the tree. Searching for 1.0401136 showed us that *HDAC1* was placed near the middle of this gene clustering. It showed up with a pair of genes, 10401109 and 10582419, listed as *Gpx2* and *Pabpn1l*, respectively. In order to evaluate the significance of this data, we needed to look at the source files from the clustering: results.cdt and results.gtr. (These are available in Appendix E.) These two files can be parsed to produce a table like the one seen in Figure 18.

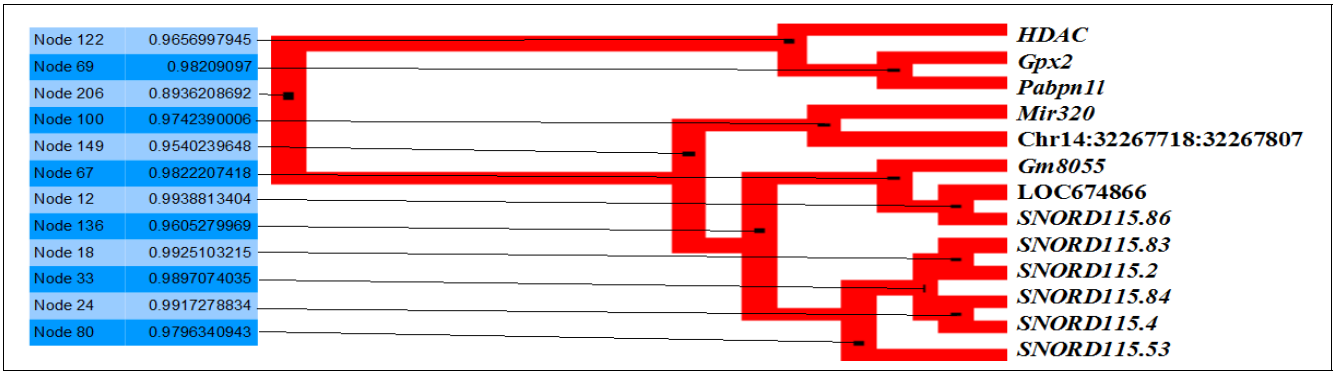

Figure 18. Greater detail of the tree highlighted in Figure 17. A table showing node values has been linked to the appropriate nodes. The gene names at each position are on the right.

Figure 18 shows how *HDAC1* and the 12 genes following it in the graph are connected. There are 12 nodes connecting the 13 genes. Each node has a correlation value showing how close they are. *Gpx2* and *Pabpn1l* are connected by Node 69, which has a correlation value of 0.98209097. There are 235 genes in the full tree, connected by 234 nodes. (For any size tree n, the number of nodes connecting them will be n-1.) The closest correlations (the highest correlation value) are built into nodes first, and receive the lower numbered labels. HDAC1 and Node 69 are joined into Node 122, with a correlation value of 0.9656997945. This was still likely significant, but tempered by the fact that over half the nodes have better correlations than Node 122. Continuing to the next node connecting 122 (and therefore *HDAC1*) and the rest of the tree, the table showed us Node 206, with a correlation of 0.8936208692. There was not enough data here to show a correlation between *HDAC1* and the other genes beyond Node 69.

In order to see a visualization of the expression levels in the third run of the data, we edited the file created by the clustering program. The trees were kept as-is, but the values were adjusted to a normalization. Each gene row had the mean and standard deviation calculated. The cell values were changed to (Original Value – GeneMean) / GeneStDev. This provided a set of data that fit Treeview's -3 to 3 range for regulation (Figure 19).

Looking at the correlation data in a heat map reinforced our earlier findings. *Gpx2* and *Pabpn1l* were very similar visually to *HDAC1*. The next two rows, corresponding to *Mir320* and *chr14:32267718:32267807*, also looked good, but the clustering sorted them away from our target gene, leaving us with not enough confidence to declare a correlation.

In order to verify the negative correlations identified in the data, we ran another test, using the raw data and a distance measure of absolute correlation. This measures positive and negative correlations and clusters accordingly (Figure 20).

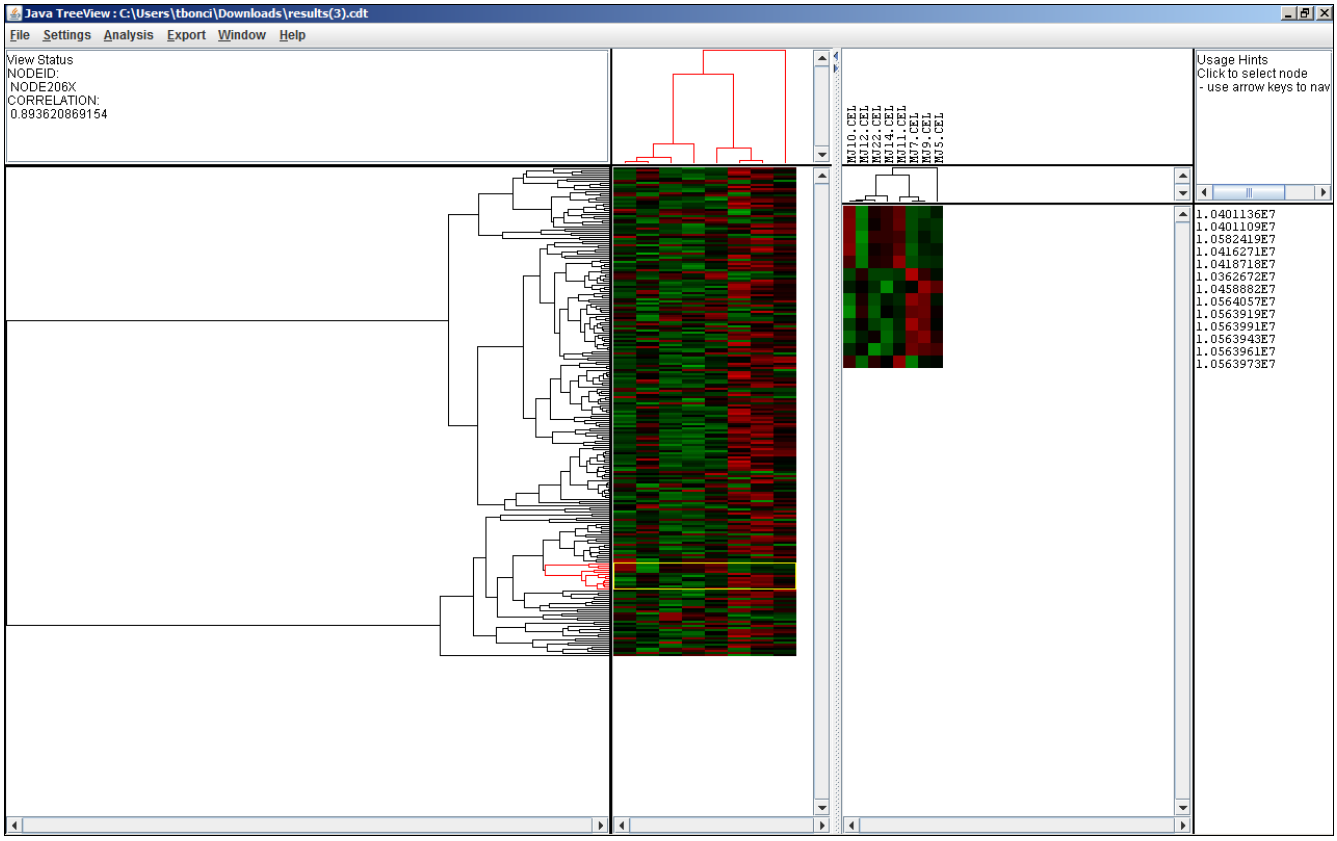

Figure 19. Treeview depiction of clustering created by correlation on raw data. The values have been normalized post-clustering to display a useful heat map. The middle panel shows the cluster selected in Figure 17.

The tree in this clustering should have in theory shown negative correlations as well as positive. The smallest cluster containing *HDAC1* also contained the genes: *Mrc1*, *Pabpn1l*, and *Npy*. The next largest cluster also contained *Gpx2*. This validated *Pabpn1l* and *Gpx2*, but we expected to see *Igk-V8-21* and *Ttr* clustered nearby as well. Unfortunately, those genes were only grouped with *HDAC1* in this clustering when nclusters < 3. In 3 or more clusters, *Igk-V8-21* and *Ttr* were separated from *HDAC1*. This test did not allow us to validate an inverse correlation between *HDAC1* and these genes. Unfortunately, the correlation coefficient for the small cluster with *Mrc1* and *Npy* was also only 0.854762980345, which was well below what we used to originally identify *Gpx2* and *Pabpn1l*.

Over the course of the four different visualizations, we were able to identify four genes of interest. The expression of *HDAC1* was shown to be positively correlated with the expression of *Gpx2* and *Pabpn1l*. The expression of *HDAC1* was also shown to be negatively (or inversely) correlated with *Igk-V8-21* and *Ttr*. These results were passed to Dr. Jakovcevski as candidates for future experiments.

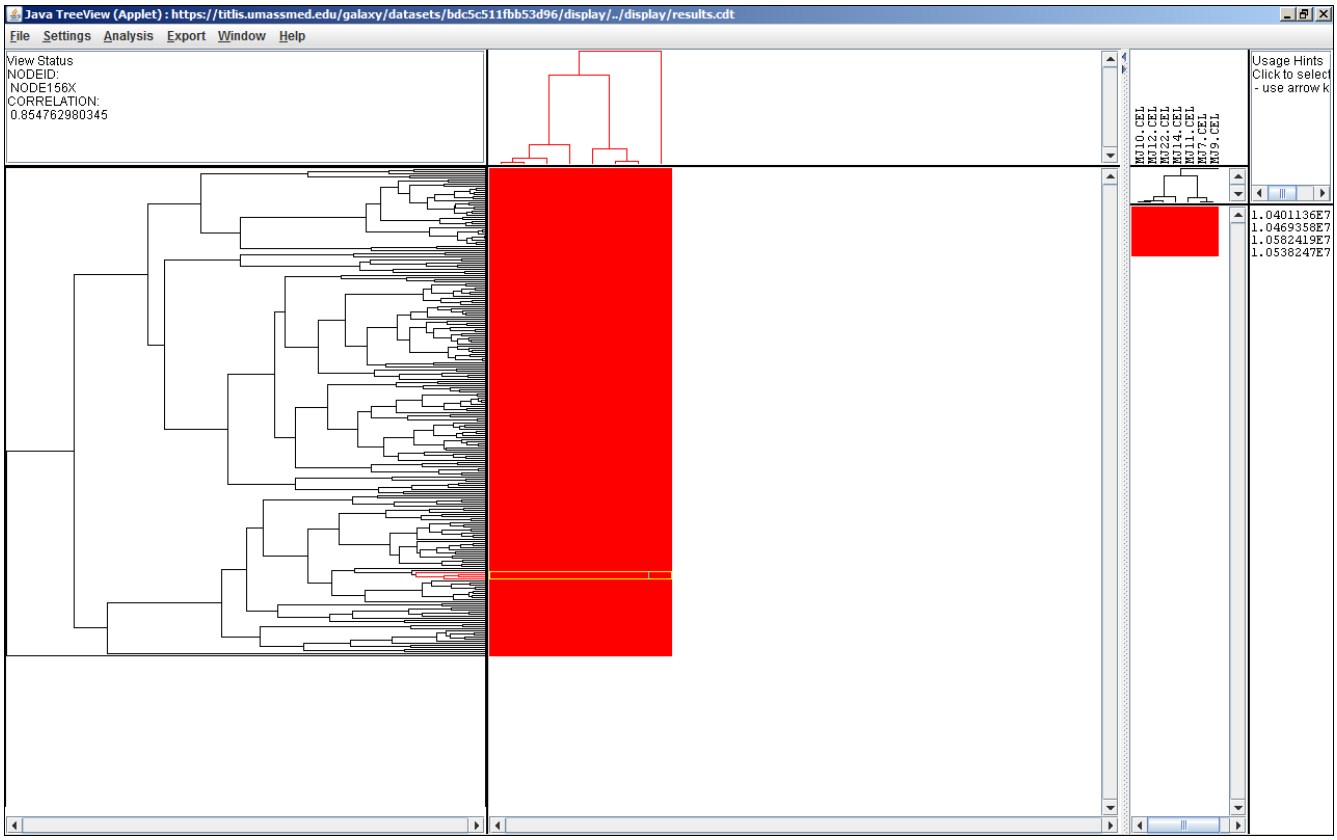

***Figure 20. Treeview visualization of raw data clustered for absolute correlation. The highlighted node shows the correlation of* HDAC1 *and the three closest genes:* Mrc1, Pabpn1l, *and* Npy.**

Discusion

Galaxy and future considerations

The integration of an applet into Galaxy will greatly expand the capabilities of the toolkit and meet a need for users requiring computational tools for gene expression data. The interactive visualization features included in Treeview are advanced beyond what is currently available from an otherwise browser-based system. The two-step process we put in place for using an applet within Treeview may become a model for other developers to follow, adding more advanced interactivity and complex functionality to the Galaxy system. In the long run, the whole system may be better implemented as a "cloud" application, similar to the ones popularly available by industry pioneers Google and Microsoft. Rebuilding the HTML interface to support a more dynamic, direct control would take a large amount of work, but the end result would be much more seamless for the user while removing the limitations of applets. As an applet, our implementation of Treeview cannot load or save presets, as it has no access to the hard drive for security reasons. A cloud based technology can interact with a local hard drive and integrate with the newest in mobile technology, where Java applets cannot.

The hierarchical clustering algorithm we made available with all four linkage types and all the different distance calculations available in the C Clustering Library. To the user, the differences between the options are not clear, and we gave no guidance in the program. Galaxy is designed with part of the tool file set aside for examples and explanations of features, essentially a "help" file. This section of the file was left with placeholder text due to time constraints. A future iteration of the clustering tool should include a full documentation of the algorithm options in the help section of the tool file.

Once the tool is iterated to a fully polished state, it can be submitted for inclusion in the universal version of Galaxy, so that every server running Galaxy will get the functionality. Currently, only the server at the University of Massachusetts Medical School has the tool included.

Data Analysis

Selecting the genes with which we worked could have been done differently. We could have clustered the full genome, but each run of the data would have taken over 45 minutes of processor time. The t-test was telling, however. A standard $p < 0.05$ left us with only a small handful of genes that showed variance. Increasing the p value to $p < 0.20$ returned ten times as many genes, enough to perform clusterings with a reasonable degree of confidence. Having the overwhelming majority of gene expression considered not statistically significant allowed us to make the decision that clustering over the whole genome would be mostly a waste of time. The set of 235 genes was felt to be large enough to calculate, while small enough to handle.

The clusterings that were run all identified the four experimental mice (MJ10, 12, 14, and 22) as similar and clustered them tightly. This is good validation of the data from these four mice. The control mice were not as perfectly clustered. MJ5 in the second and third runs was the outlier. The clustering program identified greater similarities between the other control mice (MJ7, 9, and 11) and the even-numbered mice than between those three and MJ5. The correlation numbers from the raw data show the experimental cluster with a correlation of 0.89. The MJ7, 9, 11 cluster had a correlation of 0.91. The node connecting these two groups had a correlation of 0.50. The largest node of the tree

that connected MJ5 to the 0.50 node had a correlation of 0.37. This is a very small sample size, so it's difficult to determine if there is a problem with the data, the experiment, or the execution of the experiment, or if this may be within the bounds of the larger set of acceptable data for a control mouse. It is important to contextualize these correlation numbers. This data were run off of a 235 gene sample set and not the 36k genes in the genome, which would all produce much higher correlation levels if calculated.

As HDACs act to hinder gene transcription, we expected to see a high level of HDAC1 correspond to a lower level of most everything else. We did not see this result, and the reason probably has to do with HATs. Once a nucleosome has had acetyl groups removed by action of HDAC, it is not in a permanent stage of inaccessibility. HATs (histone acetyl transferases) work in continuous cycle to undo the work of HDAC and reattach acetyl groups to histone proteins. An increase in the amount of HDAC at work did not produce a general muted expression level. Two possible explanations are: either HATs work more efficiently than HDACs, allowing "normal" levels of HATs to counteract high levels of HDACs, or our understanding of the epigenetic changes involved in gene transcription is incomplete. We know from previous works that nucleosomes without acetylation are usually associated with inactive genes, and those with acetyl groups are usually associated with active genes. This does not imply any causation, however. The proper acetylation of a nucleosome could be a directed step in targeted gene expression, as opposed to the natural/entropic action of free-floating HATs/HDACs activating as they pass near histone proteins.

The efficiency usually selected in organisms would suggest the latter to be a more reasonable explanation. If the normal level of HATs could counteract an increased number of HDACs, it's likely that equilibrium would drive an evolved organism to only produce the amount necessary for health. The counter to this argument is that HDACs in concentration are dangerous enough that a larger-than-strictly-necessary number of HATs are produced naturally to guard against the shutdown of transcription if HDACs build up in large amounts.

The two genes we did find with lower amounts are *Igk-V8-21* and *Ttr*. It's unfortunate that we were unable to validate the negative correlation through an absolute value clustering, but the clustering tool was only the first step in prompting us to look at the data for these two genes more closely. Upon doing so, the negative correlation appears probable. *Ttr* in particular is an important find because it validates a previous study showing that schizophrenia patients have abnormally low levels of TTR (transthyretin) and high levels of HDAC1 (Yang, et al., 2006). The genes with a correlated upregulation are *Gpx2* and *Pabpn1l*. *Gpx2* codes for a peroxidase, which functions to keep hydrogen peroxide from building. If the production of this peroxidase were compromised by an HDAC inhibitor, that could lead to possible toxic peroxide buildup. The next steps would be to test HDAC inhibitors against *Gpx2* and *Pabpn1l* to see if the effect on HDAC translates to a similar effect on these two genes. Also, testing HDAC inhibitors in an upregulated HDAC environment to observe TTR changes would be in order, to confirm the link between *HDAC1* and this gene.

Appendix A: Algorithms

All content taken from C Clustering Library documentation,
http://bonsai.hgc.jp/~mdehoon/software/cluster/cluster.pdf

A.1. Pearson correlation coefficient:

The Pearson correlation coefficient is defined as

$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma_x}\right)\left(\frac{y_i - \bar{y}}{\sigma_y}\right)$$

in which x, y are the sample mean of x and y respectively, and σ_x σ_y are the sample standard deviation of x and y. The Pearson correlation coefficient is a measure for how well a straight line can be fitted to a scatterplot of x and y. If all the points in the scatterplot lie on a straight line, the Pearson correlation coefficient is either +1 or −1, depending on whether the slope of line is positive or negative. If the Pearson correlation coefficient is equal to zero, there is no correlation between x and y. The *Pearson distance* is then defined as

$$d_P \equiv 1 - r\ .$$

As the Pearson correlation coefficient lies between −1 and 1, the Pearson distance lies between 0 and 2.

Note that the Pearson correlation automatically centers the data by subtracting the mean, and normalizes them by dividing by the standard deviation. While such normalization may be useful in some situations (e.g., when clustering gene expression levels directly instead of gene expression ratios), information is being lost in this step. In particular, the magnitude of changes in gene expression is being ignored. This is in fact the reason that the Pearson distance does not satisfy the triangle inequality.

A.2. Absolute Pearson correlation:

By taking the absolute value of the Pearson correlation, we find a number between zero and one. If the absolute value is one, all the points in the scatter plot lie on a straight line with either a positive or a negative slope. If the absolute value is equal to zero, there is no correlation between x and y.

The distance is defined as usual as

$$d_A \equiv 1 - |r|\ ,$$

where r is the Pearson correlation coefficient. As the absolute value of the Pearson correlation

coefficient lies between 0 and 1, the corresponding distance lies between 0 and 1 as well.

In the context of gene expression experiments, note that the absolute correlation is equal to one if the gene expression data of two genes/microarrays have a shape that is either exactly the same or exactly opposite. The absolute correlation coefficient should therefore be used with care.

A.3. Uncentered correlation (cosine of the angle):

In some cases, it may be preferable to use the *uncentered correlation* instead of the regular Pearson correlation coefficient. The uncentered correlation is defined as

$$r_U = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i}{\sigma_x^{(0)}}\right)\left(\frac{y_i}{\sigma_y^{(0)}}\right),$$

where

$$\sigma_x^{(0)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2};$$
$$\sigma_y^{(0)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^2}.$$

This is the same expression as for the regular Pearson correlation coefficient, except that the sample means *x*, *y* are set equal to zero. The uncentered correlation may be appropriate if there is a zero reference state. For instance, in the case of gene expression data given in terms of log-ratios, a log-ratio equal to zero corresponds to the green and red signal being equal, which means that the experimental manipulation did not affect the gene expression.

The distance corresponding to the uncentered correlation coefficient is defined as

$$d_U \equiv 1 - r_U,$$

where r_U is the uncentered correlation. As the uncentered correlation coefficient lies between −1 and 1, the corresponding distance lies between 0 and 2.

The uncentered correlation is equal to the cosine of the angle of the two data vectors in *n*-dimensional space, and is often referred to as such. (From this viewpoint, it would make more sense to define the distance as the arc cosine of the uncentered correlation coefficient).

A.4. Absolute uncentered correlation:

As for the regular Pearson correlation, we can define a distance measure using the absolute value of the uncentered correlation:

$$d_{AU} \equiv 1 - |r_U| \ ,$$

where r_U is the uncentered correlation coefficient. As the absolute value of the uncentered correlation coefficient lies between 0 and 1, the corresponding distance lies between 0 and 1 as well.

Geometrically, the absolute value of the uncentered correlation is equal to the cosine between the supporting lines of the two data vectors (i.e., the angle without taking the direction of the vectors into consideration).

A.5. Spearman rank correlation:

The Spearman rank correlation is an example of a non-parametric similarity measure. It is useful because it is more robust against outliers than the Pearson correlation.

To calculate the Spearman rank correlation, we replace each data value by their rank if we would order the data in each vector by their value. We then calculate the Pearson correlation between the two rank vectors instead of the data vectors.

Weights cannot be suitably applied to the data if the Spearman rank correlation is used, especially since the weights are not necessarily integers. The calculation of the Spearman rank correlation in the C Clustering Library therefore does not take any weights into consideration.

As in the case of the Pearson correlation, we can define a distance measure corresponding to the Spearman rank correlation as

$$d_S \equiv 1 - r_S,$$

where r_S is the Spearman rank correlation.

A.6. Kendall's τ:

Kendall's τ is another example of a non-parametric similarity measure. It is similar to the Spearman rank correlation, but instead of the ranks themselves only the relative ranks are used to calculate τ (see Snedecor & Cochran). As in the case of the Spearman rank correlation, the weights are ignored in the calculation.

We can define a distance measure corresponding to Kendall's τ as

$$d_K \equiv 1 - \tau \ .$$

As Kendall's τ is defined such that it will lie between −1 and 1, the corresponding distance will be between 0 and 2.

A.7. Euclidean distance:

The Euclidean distance is a true metric, as it satisfies the triangle inequality. In this software package, we define the Euclidean distance as

$$d = \frac{1}{n} \sum_{i=1}^{n} (x_i - y_i)^2 .$$

Only those terms are included in the summation for which both x_i and y_i are present. The denominator n is chosen accordingly.

In this formula, the expression data x_i and y_i are subtracted directly from each other. We should therefore make sure that the expression data are properly normalized when using the Euclidean distance, for example by converting the measured gene expression levels to log-ratios.

Unlike the correlation-based distance functions, the Euclidean distance takes the magnitude of the expression data into account. It therefore preserves more information about the data and may be preferable. De Hoon, Imoto, Miyano (2002) give an example of the use of the Euclidean distance for *k*-means clustering.

A.8. City-block distance:

The city-block distance, alternatively known as the Manhattan distance, is related to the Euclidean distance. Whereas the Euclidean distance corresponds to the length of the shortest path between two points, the city-block distance is the sum of distances along each dimension. As gene expression data tend to have missing values, in the C Clustering Library we define the city-block distance as the sum of distances divided by the number of dimensions:

$$d = \frac{1}{n} \sum_{i=1}^{n} |x_i - y_i| .$$

This is equal to the distance you would have to walk between two points in a city, where you have to walk along city blocks. The city-block distance is a metric, as it satisfies the triangle inequality. As for the Euclidean distance, the expression data are subtracted directly from each other, and we should therefore make sure that they are properly normalized.

Appendix B: Source code

B.1. prep4jtv.py:

```
"""
Usage: prep4jtv.py
        infile
        outfile1.xxx
        gmethod(s=single-linkage m=maximum c=centroid-linkage
a=average-linkage)
        gdist(c=correlation a=abs(correlation) u=uncentered correlation
x=abs(uncentered correlation) s=Spearman's k=Kendall's e=Euclidian
b=City-block)
        bool cluster_expression
        emethod
        edist

Author: Tim Bonci
"""


from Pycluster import *
import sys
import os
import os.path

handle = open(sys.argv[1])
rec = Record(handle)

path = sys.argv[2]
(fname,fext) = os.path.splitext(sys.argv[2])
fname = fname+"_files"
os.mkdir(fname)
fname = fname+"/results"

genetree = rec.treecluster(method=sys.argv[3], dist=sys.argv[4])
genetree.scale()
if bool(sys.argv[5])==True:
        exptree = rec.treecluster(method=sys.argv[6], dist=sys.argv[7],
transpose=1)
```

```
        rec.save("cluster", genetree, exptree)
        os.system("cp cluster.atr "+fname+'.atr')
else:
        rec.save("cluster", genetree)

os.system("cp cluster.cdt "+fname+'.cdt')
os.system("cp cluster.gtr "+fname+'.gtr')

rval = '''<html><head><title><Composite Dataset (cdt)</title></head>
        <body>
        <APPLET
CODE=edu/stanford/genetics/treeview/applet/ButtonApplet.class
archive=../../../static/treeview/TreeViewApplet.jar,../../../static/treeview/nanoxml-
2.2.2.jar,../../../static/treeview/plugins/Dendrogram.jar
width=150
hright=50
alt="Your browser understands the &lt;APPLET&gt; tag but is not running
the applet.">
<param name="cdtFile" value="../display/results.cdt">
<param name="cdtName" value="results.cdt">
<param name="plugins"
value="edu.stanford.genetics.treeview.plugin.dendroview.DendrogramFactory">
Your browser is ignoring the &lt;APPLET&gt; tag.
</APPLET>
</BODY></HTML>'''
fout = open(sys.argv[2], "w")
fout.write(rval)
fout.close
```

B.2. prep4jtv.xml:

```
<tool id="prep4jtv" name="Java TreeView" version="0.2.0">
  <description>Applet to visualize clusters</description>
  <command interpreter="python">prep4jtv.py $infile
$outfile $method $dist $xcluster $xmethod $xdist
  </command>
  <inputs>
    <param name="infile" type="data" label="Input File">
    </param>
```

```
    <param name="method" type="select" label="Clustering Method">
      <option value="s">Pairwise single-linkage</option>

      <option value="m">Pairwise maximum- (or complete-)
linkage</option>
      <option value="c">Pairwise centroid-linkage</option>
      <option value="a">Pairwise average-linkage</option>
    </param>
    <param name="dist" type="select" label="Distance Function">
      <option value="c">Correlation</option>
      <option value="a">Absolute value of the correlation</option>

      <option value="u">Uncentered correlation</option>
      <option value="x">Absolute uncentered correlation</option>
      <option value="s">Spearman's rank correlation</option>
      <option value="k">Kendall's</option>
      <option value="e">Euclidian distance</option>
      <option value="b">City-block distance</option>

    </param>
    <param name="xcluster" type="boolean" label="Cluster Expression
Conditions" value="FALSE">
    </param>
    <param name="xmethod" type="select" label="Expression Clustering
Method">
      <option value="s">Pairwise single-linkage</option>
      <option value="m">Pairwise maximum- (or complete-)
linkage</option>
      <option value="c">Pairwise centroid-linkage</option>

      <option value="a">Pairwise average-linkage</option>
    </param>
    <param name="xdist" type="select" label="Expression Distance
Function">
      <option value="c">Correlation</option>
      <option value="a">Absolute value of the correlation</option>
      <option value="u">Uncentered correlation</option>
      <option value="x">Absolute uncentered correlation</option>
```

```
      <option value="s">Spearman's rank correlation</option>
      <option value="k">Kendall's</option>
      <option value="e">Euclidian distance</option>
      <option value="b">City-block distance</option>
    </param>
  </inputs>
  <outputs>

    <data format="html" name="outfile"
label="Clustering of $infile.name"/>
  </outputs>
  <help>
No help written yet.
  </help>
</tool>
```

Appendix C: Gene Expression Data, p < 0.2

C.1 Microarray Data:

	MJ10	MJ11	MJ12	MJ14	MJ22	MJ5	MJ7	MJ9
10346168	143.2	155.46	129.82	107.42	129.73	189.36	204.75	173.22
10346235	131.38	190.3	131.58	140.12	137.74	189.13	161.35	150.42
10346551	144.61	165.75	138.74	147.27	149.23	217.81	166.89	168.88
10348299	123.29	154.49	126.04	118.16	131.39	168.71	161.95	126.58
10349174	295.36	364.41	293.56	286.32	294.31	334.14	387.08	347.24
10349947	220.13	281.67	251.75	199.49	228.83	314.85	295.01	256.7
10351041	249.96	220.09	217.34	235.45	262.17	185.32	213.37	179.74
10354414	160.8	122.3	171.73	164.02	166.09	128.63	146.93	150.02
10356145	201.24	258.31	205.05	219.23	203.44	246.24	270.83	228.78
10358339	226.62	290.84	207.85	228.59	239.4	418.04	287.29	275.18
10360391	178.89	155.78	132.01	97.5	195.57	275.89	400.78	335.26
10361897	656.02	662.17	582.81	545.7	705.45	843.45	768.34	735.88
10362450	173.6	128.17	184.13	152.44	169.6	121.51	156.09	154.44
10362672	2449.39	1850.46	2273.3	2248.99	2517.97	1873.01	2080.6	2026.13
10363007	396.86	497.43	421.43	411.81	382.93	503.41	465.55	462.44
10363157	126.93	149.6	130.28	124.49	104.8	153.33	160.89	151.18
10364593	159.17	180.29	156.03	155.82	171.15	213.5	215.96	198.18
10365974	446.71	544.53	393.83	451.52	443.25	886.05	547.18	563.08
10366476	330.55	367.67	340.39	349.77	336.92	450.68	478.3	385.2
10368289	313.62	327.86	271.14	312.44	265.87	450.92	387.95	380.91
10368370	183.86	134.51	159.9	160.16	149.27	120.93	136.99	146.45
10370327	258.48	202.52	258.53	245.32	253.4	166.42	222.39	216.75
10373358	104.75	94.44	125.17	117.35	112.72	79.75	105.65	89.84
10373542	463.57	593.59	466.37	478.08	505.86	593.18	653.58	611.49
10373606	99.07	92.05	116.62	138.14	128.51	95.27	91.94	106.79
10373610	688.66	498.91	802.24	699.58	883.64	495.25	622.95	636.33
10375129	183.08	148.75	199.68	195.39	193.73	151.86	162.64	184.77

10375443	250.71	283.53	260.51	198.05	325.75	339.71	401.04	382.89
10376060	199.09	199.28	185.19	135.64	199.4	239.5	352.92	299.14
10376163	363.68	450.24	351.9	328.07	329.14	430.28	369.56	400.98
10376324	322.65	335.92	230.84	169.33	331.25	398.56	894.59	683.32
10376326	453.88	489.35	409.91	225.08	584.81	692.55	1306.4	1095.09
10376532	142.24	115.01	150.92	161.41	152.34	116.31	132.55	136.4
10376885	2260.67	1714.11	1654.46	1683.42	1777.8	1371.9	1186.29	1266.79
10379190	744.37	955.19	745.91	832.67	859.31	1053.38	1196.16	893.45
10379389	301.91	304.66	281.8	269.01	341.01	369.71	386.62	375.65
10379953	207	239.05	193.29	204.08	214.08	255.21	268.3	253.87
10381334	131.17	85.08	127.72	112.8	115.05	102.55	101.39	104.38
10384725	320.77	342.64	283.29	297.7	326.67	388.37	399.25	356.87
10385504	93.21	98.38	78.59	86.44	99.58	122.99	174.03	116
10385511	271.47	281.49	249.57	182.22	298.29	352.79	433.88	370.95
10387219	194.34	134.8	215.09	218.3	189.03	180.46	171.83	185.56
10388430	310.27	367.61	314.81	311.3	346.08	467.41	368.16	371.78
10389025	371.87	354.33	296.65	345.51	352.23	477.2	413.95	413.39
10390931	1309.67	947.74	1405.75	1335.98	1518.72	957.57	1157.38	1316.07
10392221	137.8	187.99	159.92	164.93	173.33	204.95	208.04	165.23
10392440	196.16	202.58	188.34	170.42	204.98	301.46	256.65	224.77
10395039	510.33	563.37	488.39	392.59	525.96	618.75	743.77	605.76
10396608	195.52	214.71	204.63	184.88	214.34	257.81	239.44	239.91
10397428	341.99	422.04	319.69	333.95	315.47	382.05	393.79	377.15
10399581	175.98	154.7	168.93	213.11	197.59	141	135.09	168.98
10401109	217.76	187.44	221.48	193.75	238.71	166.19	186.47	196.66
10401136	4624.84	326.82	4248.59	1568.15	6980.64	388.32	388.82	333.41
10403246	460.34	361.39	469.82	434.41	493.78	344.65	398.95	417.5
10404389	110.04	141.06	102.01	92.89	107.77	105.06	186.03	144.08
10404429	136.98	164.56	123.72	139.93	139.97	187.64	191.17	170.85
10405587	289.06	319.04	282.02	311.63	332.8	478.8	423.13	362.46
10405785	165.92	177.14	156.57	154.7	137.97	205.88	175.61	172.12

10407173	1030.51	1084.93	974.05	941.46	1171.58	1478.92	1348.31	1207.92
10408741	321.51	401.65	315.22	290.44	305.11	388.56	358.68	357.04
10408870	361.22	518.29	315.22	441.79	363.59	504.94	429	440.19
10412298	178.59	211.73	177.07	166.36	206.89	271.82	275.3	228.78
10413803	232.32	282.15	222.77	237.69	232.42	324.07	255.43	265.19
10414706	573.63	406.51	600.11	570.96	662.52	449.55	524.02	576.78
10415952	120.9	116.12	94.95	111.86	116.8	146.62	136.35	155.45
10416271	542.27	413.37	527.52	563.04	557.81	445.24	481.28	483.58
10416657	246.87	276.76	258.47	247.56	292.28	348.28	375.8	297.1
10418718	165.69	121.11	158.12	167.92	172.12	131.99	128.72	136.94
10419900	121.11	98.8	130.8	129.58	131	95.97	112.73	116.92
10420308	90.04	98.63	90.67	76.75	87.56	99.3	165.18	122.21
10420899	151.24	127.65	147.56	157.98	173.09	120.1	138.94	118.91
10423654	143.3	108.34	147.62	133.77	154.9	114.78	126.05	135.46
10424287	121.64	101.03	140.23	134.05	138.9	109.37	107.18	118.57
10424695	285.91	205.24	308.73	254.04	264.38	225.84	226.07	243.66
10426098	581.51	679.36	547.45	566.3	584.88	747.17	774.65	645.05
10428509	275.55	413.22	266.76	344.6	236.05	321.4	326.5	371.88
10429638	242.88	207.06	242.07	210.78	265.69	143.18	214.15	171.64
10430645	86.41	114.65	104.2	87.87	95.66	127.51	115.1	128.44
10430899	292.22	215.18	291.56	275.67	320	235.09	273.83	270.82
10433101	268.91	280.32	238.94	204.49	281.6	341.64	327.61	302.24
10433431	199.67	233.17	191.35	212.8	188.86	254.12	288.38	239.37
10435266	195.81	176.23	167.68	178.05	173.82	252.42	249.27	231.12
10435767	265.86	442.55	298.81	297.3	241.74	429.66	309.56	382.26
10436200	289.73	235.58	335.71	288.1	326.67	237.68	271.48	288.48
10436500	248.69	280.67	244.79	251.19	251.35	384.99	309.06	265.09
10436590	129.68	134.48	105.21	126.44	120.56	172.3	138.76	137.87
10436658	251.16	350.49	277.53	271.9	248.14	275.43	299.3	330.3
10436830	546.53	573.89	461.97	470.62	609.93	780.37	701.33	627.38
10436841	311.68	343.32	307.96	275.54	361.6	427.16	446.4	371.83

10436890	442.85	354.55	368.24	374.8	471.33	302.78	335.9	271.24
10438517	866.43	1022.5	827.88	1022.08	880.22	1240.22	1118.45	954.38
10444229	385.99	423.32	306.49	268	396.94	415.68	550.27	608.93
10444291	354.29	712.91	372.51	171.04	830.15	935.07	2377.2	2364.93
10444298	377.41	479.37	343.67	323.6	491.95	595.42	1048.12	1267.53
10445909	211.55	223.44	186.94	213.68	217.2	293.74	263.72	235.21
10447097	125.98	149.92	142.11	106.11	129.18	157.88	139.01	152.04
10450154	395.29	714.52	336.14	179.85	791.62	898.92	2079.73	2523.68
10450191	192.67	156.63	188.42	223.86	217.96	176.37	176.61	142.51
10452404	251.21	286.91	221.32	252.23	232.79	329.99	280.28	248.46
10453057	130.58	148.31	134.48	127.91	146.94	185.6	163	142.39
10454192	163.85	1888.13	174.71	196.89	279.48	11677.7	1260.79	239.73
10454310	203.73	232.44	208.47	208.83	223.31	276.87	259.14	245.77
10454828	239.94	328.66	223.38	251.1	229.77	275.43	275.42	283.56
10455118	193.12	244.97	172.94	196.02	177.2	244.01	244.18	210.91
10456005	467.39	837.14	430.15	154.6	1092.54	1134.62	3070.08	2978.31
10458314	305.12	325.65	287.76	275.1	378.19	431.75	491.6	406.07
10458882	280.88	209.27	258.2	264.71	300.9	229.5	230.71	229.88
10459618	388.52	313.73	433.04	385.79	348.45	287.77	315.08	313.89
10461012	229.21	322.08	227.94	245.95	235.27	303.85	261.48	245.08
10461115	383.9	472.03	397.42	409.71	387.52	701.32	553.64	408.95
10461156	494.59	378	385.61	416.46	412.94	350.09	332.22	301.27
10461723	114.86	141.44	123.04	98.23	129.67	147.08	150.43	135.88
10463070	457.78	514.39	463.1	407.89	598.72	635.73	718.26	659.65
10466288	137.25	110.38	168.64	163.55	145.56	132.73	131.74	136.23
10466521	150.52	156.06	154.65	125.86	162.02	211.93	182.22	180.68
10466938	203.29	254.25	197.05	193.93	190.77	219	234.42	229.17
10467110	217.11	229.06	223.85	208.09	207.03	298.3	248.12	267.48
10469151	193.69	251.57	227.96	230.94	225.45	329.54	292.24	233.18
10469255	212.53	232.52	188.58	209.66	223.73	375.28	255.09	271.53
10469358	104.47	137.12	115.28	127.3	108.75	147.36	141.64	139.05

10471486	358.94	416.41	380.22	369.3	399.48	549.91	538.44	413.27
10475199	962.63	1005.19	953.18	946.31	1127.69	1374.11	1248.91	1192.91
10477920	207.44	241.15	215.65	224.4	236.98	285.6	298.76	243.12
10479971	277.2	333.29	289.84	236.53	265.67	306.8	319.2	351.05
10479975	323.17	421.74	345.84	377.36	332.8	380.77	439.05	416.2
10483046	136.59	143.52	136.49	114.84	118.46	157.04	164.15	143.02
10484203	275.75	343.25	237.04	262.09	302.44	443.48	415.92	323.7
10484207	286.57	308.03	258.81	263.87	239.99	402.35	342.67	302.85
10484463	407.03	420.04	301.91	235.06	558.16	783.36	1149.13	732.53
10486396	255.59	310.4	237.35	230.62	285.15	312.87	331.07	283.49
10487252	249	317.8	248.09	247.67	233.25	319.27	260.23	279.85
10488912	282.55	320.44	262.97	246.91	288.64	358.93	372.16	319.42
10489107	521.55	522.12	441.01	355.85	565.93	720.6	1076.3	763.87
10490872	142.93	182.24	128.33	145.7	144.3	172.21	161.73	167.43
10490989	288.76	344.71	288.1	253.99	363.3	580.04	481.69	438.03
10491091	182.09	199.75	171.17	150.2	188.23	230.03	347.12	274.32
10491846	103.59	120.73	111.06	91.51	87.27	110.23	124.85	125.92
10492558	121.59	114.95	97.41	101.98	111.4	140.51	140.04	140.9
10494978	89.36	98.15	91.77	82.86	102.25	116.83	180.23	162.27
10495054	258.05	298.87	242.3	236.85	294.03	390.89	308.17	295.92
10495685	178.28	239.65	176.17	173.08	169.31	210.98	199.6	190.37
10496359	314.89	412.09	325.19	347.34	334.88	671.54	408.82	397.54
10496872	315.73	398.72	329.54	374.31	350.38	539.06	519.26	397.18
10497773	365.33	409.83	304.51	352.17	324.97	478.69	394.12	362.47
10497971	135.1	181.96	143.44	141.34	141.95	181.75	165.23	153.38
10498350	137.72	169.68	163.2	163.21	174.62	211.97	222.98	183.96
10499093	152.7	128.89	154.63	151.57	145.01	119.11	102.41	126.42
10499189	182.6	187.84	203.06	197.32	273.15	167.34	158.84	134.78
10500677	109.92	113.34	110.44	117.62	111.74	138.72	155.5	141.03
10502201	136.67	127.06	128.08	137.51	143.47	100.05	123.28	108.33
10505270	966.56	1027.44	933.44	895.87	1051.84	1154.35	1270.25	1185.6

10505954	236.27	269.79	248.31	227.92	244.84	306.65	365.56	254.24
10510872	357.4	262.62	381.93	334.16	365.57	282.33	324.65	334.21
10514323	245.94	182.41	258.59	229.33	271.42	185.19	237.07	219.78
10514398	147	171.41	129.18	129.12	145.3	181.65	156.63	153.39
10517573	174	141.25	176.68	214.87	179.21	152.19	164.34	151.83
10518408	423.32	446.58	369.2	409.74	449.7	579.6	545.03	473.56
10519527	346.44	380.64	313.71	331.53	341.23	453.21	485.8	391.57
10519555	92.94	129.17	95.36	96.87	128.71	136.01	148.7	116.65
10519607	746.71	803.35	706.3	745.53	899.29	1020.58	1068.78	956.48
10520445	155.44	127.08	178.67	182.95	187.2	154.01	112.54	159.84
10525439	328.74	364.57	337.52	309.6	363.75	441.16	429.18	406.83
10527649	278.34	318.62	226.02	245.06	247.78	286.3	312.78	272.2
10527963	116.82	152.13	102.59	106.17	89.94	112.53	121.19	132.01
10528622	167.42	147.88	190.48	187.31	181.69	132.76	160.7	153.13
10530492	110.63	148.87	123.39	131.37	150.14	154.68	162.08	176.01
10530841	817.39	1032.13	820.64	858.1	980.72	1955.86	1531.92	1025.64
10531407	448.8	548	419.58	234.85	592.87	621.22	1183.96	1385.67
10531987	228.45	272.16	197.74	117.96	292.46	381.16	1213.04	719.68
10534301	221.18	189.17	218.49	208.86	217.79	171.47	181.07	167.96
10535807	374.72	429.17	363.3	374.35	398.3	503.76	518.25	410.69
10537227	140.72	170.42	159.98	133.31	157.79	182.04	296.71	275.91
10538247	1925.64	2404.98	1924.04	2118.41	1950.01	2493.73	2340.73	2405.3
10540105	730.41	887.32	637.89	692.65	690.01	901.54	892.48	761.57
10540493	435.34	506.35	392.02	431.15	541.86	587.43	737.49	599.71
10541318	290.42	354.7	292.82	234.72	260.38	336.52	312.52	306.57
10544523	1357.66	983.7	1135.14	1369.86	1278.35	767.41	750.96	1008.72
10544588	99.85	97.48	77.23	76.04	93.12	125.4	157.88	134.85
10545239	80.1	5108.42	82.99	69.46	204.48	127.61	7431.19	7322.11
10548892	337.55	339.63	303.88	285.64	407.62	558.83	559.71	515.75
10548996	518.19	662.04	489.52	586.96	543.66	690.85	645.39	598.53
10549377	127.4	105.07	117.99	118.58	118	94.85	103.92	89.94

10552824	970.15	1017.72	881.9	975.85	1075.48	1344.02	1359.73	1159.82
10555323	229.47	249.46	195.04	209.32	237.78	261.71	285.49	244.16
10560190	150.02	183.25	167.69	170.87	157.84	217.11	199.7	192.57
10560242	107.38	120.84	116.64	98.24	134.91	146.79	156.04	143.34
10562166	316.25	212.92	291.96	308.14	240.79	211.24	174.34	254.71
10562192	1029.86	1178.53	1092.08	981.79	1155.7	1188.58	1465.09	1476.01
10563919	2360.19	1293.98	1873.34	2049.29	2148.02	1564.08	1690.41	1734.61
10563933	447.71	271.54	322.9	381.39	371.59	286.98	309.94	331.51
10563943	3480.32	2328.01	3044.16	3149.41	3363.88	2551.78	2613.38	2785.47
10563961	834.93	389.09	627.74	693.4	787.21	461.58	528.52	509.66
10563973	2438.16	1434.42	2153.8	2156.23	2212.04	1692.04	1897.42	1790.81
10563991	1562.6	859.8	1241.51	1290.22	1436.97	947.61	1114.12	1136.85
10564057	1903.11	1069.94	1728.13	1764.41	2226.95	1255.19	1383.66	1419.58
10565018	287.53	353.29	287.97	289.1	298.03	442.83	362.3	320.73
10566574	437.02	470.37	420.59	319.69	442.2	459.98	622.39	560.61
10572432	469.62	578.05	494.88	539.02	525.62	667.88	640.5	541.47
10573578	1322.56	1863.98	1462	1727.72	1658.33	2025.6	2053.82	1768.77
10574438	200.48	230.91	199.63	186.22	213.18	297.56	268.53	225.61
10576049	135.65	153.77	163.63	134.04	150.48	185.62	193.87	168.17
10576661	433.54	463.57	367.53	404.8	451.81	531.43	515.16	503.71
10578685	204.85	163.13	199.34	188.71	226.79	156.55	185.11	187.2
10579894	214.67	201.25	182.05	155.93	187.76	251.29	215.59	279.17
10580033	192.41	212.54	182.95	187.63	205.26	286.81	295.09	252.33
10581378	505.31	619.74	530.21	452.07	558.7	597.04	657.62	604.5
10581388	573.07	656.7	565.52	567.93	602.14	751.09	775.61	666.2
10582419	196.25	153.73	188.49	169.34	206.82	140.79	157	158.27
10583145	408.44	445.84	346.54	395.92	517.97	716.3	740.55	567.39
10583203	175.37	175.24	130.68	142.29	165.18	285.1	260.67	171.46
10584589	96.26	119.29	83.22	94.63	106.94	126.55	118.19	114.04
10585697	371.66	300.17	315.41	381.49	339.12	263.74	285.99	306.42
10586172	111.14	124.38	103.75	122.34	111.14	141.73	142.6	127.98

10586441	271.56	362.77	269.42	302.71	269.3	355.51	303.39	323.53
10588079	251.87	186.28	276.94	247.29	240.64	210.7	226.08	219
10588201	194.06	175.01	227.29	211.39	187.03	139.01	167.76	163.27
10588849	201.2	179.31	224.3	202.31	210.34	183.18	164.61	178.59
10588931	136.31	118.26	154.75	133.38	148.66	117.05	121.7	114.55
10590860	245.34	319.76	276.1	233.74	261.63	335.82	284.15	283.68
10592449	180.55	158.7	199.35	202.83	209.71	156.55	191.76	158.37
10593050	258.81	268.78	250.82	211.54	296.63	318.28	398.17	358.52
10595439	277.52	254.99	317.51	310.73	313.54	233.45	257.72	277.8
10595768	271.52	269.67	223.74	255.5	237.39	317.65	301.23	291.6
10597518	349.01	374.58	375.85	340.25	410.53	491.8	468.59	465.41
10597648	266.64	289.95	235.24	231.69	298.78	360.41	380.41	307.06
10597960	635.66	684.12	587.65	563.64	600.26	988.22	741.21	661.58
10598075	1403.53	913.4	971.38	1357.73	1140.45	820.57	746.95	834.94
10600169	403.53	428.64	344.43	417.22	398.93	612.61	552.01	458.41
10600584	121.76	100.16	151.06	121.8	118.19	112.65	114.27	100.59
10604832	372.85	396.81	320.9	359.93	341.42	484.21	474.78	445.65
10606301	297.82	307.8	268.79	280.38	319.03	412.67	347	337.09
10606389	155.81	117.89	163.35	177.81	162.49	128.08	150.63	149.6
10606542	2698.22	2134.4	2924.01	2928.85	3298.67	2265.6	2473.29	2466.02
10607116	159.27	224.87	146.47	167.38	156.35	218.18	186.32	167.13
10608136	370.66	307.67	339.14	370.45	388.85	273.63	297.65	338.9

C.2. Microarray labels:

10346168,Stat4
10346235,Hibch
10346551,Cflar
10347158,ENSMUST00000119302 /// GENSCAN00000023709
10348299,5830472F04Rik
10348996,GENSCAN00000039359
10349174,Serpinb8
10349947,Fmod
10351041,ENSMUST00000083801
10351045,FM991906

10351206,Selp
10354414,chr1:49635994:49636346:-
10356145,Slc19a3
10357090,Serpinb3c
10358339,Cfh
10358553,Hmcn1
10358561,Hmcn1
10358587,Hmcn1
10359504,Dnm3os
10360391,Ifi203
10361897,Ifngr1
10362350,Themis
10362450,Trdn
10362672,Gm8055
10363007,Ascc3
10363157,Pln
10364593,Cnn2
10365974,Dcn
10366476,Ptprb
10366512,ENSMUST00000083152
10367473,Olfr763
10367517,Olfr804
10368289,Enpp1
10368370,Gm8681
10368947,Aim1
10370327,chr10:77197596:77197890:-
10373358,Il23a
10373542,Dgka
10373606,Olfr765
10373610,Olfr767
10375129,D130052B06Rik
10375443,Havcr2
10376060,Irf1
10376163,Rapgef6
10376324,Gm12250
10376326,Irgm2 /// Igtp
10376532,Olfr225
10376885,Snord49b
10379190,Vtn
10379389,Adap2
10379953,4632419I22Rik
10381334,Cntd1 /// Becn1
10384725,Rel
10384770,5730522E02Rik

10385504,Gm5431
10385511,Psme2 /// Psme2b-ps
10386622,LOC100047860 /// LOC677463
10387219,Rangrf /// Slc25a35
10388430,Serpinf1
10389025,Myo1d
10390931,Krtap4-7
10392221,Pecam1
10392440,Slc16a6
10395039,Cmpk2
10396608,Syne2
10397428,1700020O03Rik
10398193,3110018I06Rik
10399581,3110053B16Rik /// LOC100046614
10401109,Gpx2
10401136,Gm4864 [HDAC1]
10403023,Gm7016
10403046,AI324046
10403073,Ighg
10403246,Gm8598
10403871,Aoah
10404389,Irf4
10404429,Serpinb9
10405587,Tgfbi
10405785,0610007P08Rik
10405885,ENSMUST00000099414
10407173,Il6st
10407995,Olfr1370
10408741,Txndc5
10408870,Tbc1d7
10410375,Zfp85-rs1
10410919,Gm5666
10412298,Itga1
10413803,Btd
10414470,Tlr11
10414706,AJ311366 /// ENSMUST00000103569
10415952,GENSCAN00000010976
10416271,mmu-mir-320
10416657,Elf1
10418718,chr14:32267718:32267807:-
10419900,Myh6 /// Myh7
10420247,Mcpt4
10420308,Gzmb
10420899,Gulo

10421517,Cysltr2
10423654,Osr2
10424287,Fer1l6
10424695,Gpihbp1
10426098,Creld2
10427669,GENSCAN00000024006
10428509,Csmd3
10429638,Gm9568
10430645,ENSMUST00000100430
10430899,Cyp2d40
10433101,Gpr84
10433431,chr16:6526419:6526726:+
10435266,Heg1
10435767,AK134586 /// ENSMUST00000099760
10436200,Gm8824
10436500,Gbe1
10436590,2810055G20Rik
10436658,7120432I05Rik
10436830,Ifnar2
10436841,Il10rb
10436890,Gm10785
10438517,Alg3
10439296,Stfa2
10440700,AY026312
10442139,Gm7673
10442149,V1re1
10444229,H2-DMa
10444291,H2-Ab1
10444298,H2-Eb1
10445176,Olfr134
10445909,Kat2b
10447079,4921513D11Rik
10447097,Gemin6
10450154,H2-Aa
10450191,Btnl5
10450603,ENSMUST00000083840
10450866,Olfr97
10450874,Olfr101
10452404,Nudt12
10453057,Cyp1b1 /// 1700038P13Rik
10453867,Rbbp8
10454192,Ttr
10454310,Galnt1
10454828,Pnet-ps

10455118,Pcdhb18
10456005,Cd74
10457918,Gm5064
10458314,Tmem173
10458882,LOC674866
10459618,ENSMUST00000082772
10461012,Trmt112 /// Prdx5
10461115,Slc22a8
10461156,Snhg1
10461723,Fam111a
10463070,Entpd1
10465224,Gm10815
10466172,Ms4a1
10466288,Olfr1428
10466521,Gcnt1
10466606,Anxa1
10466938,5033414D02Rik
10467110,AI747699
10469151,Itih5
10469255,Prkcq
10469278,Il2ra
10469358,Mrc1
10470141,Lcn8
10471486,Eng
10471780,Olfr367
10473608,Olfr1193
10475199,Snap23
10475910,chr2:128606112:128606235:+
10477920,Myl9
10478523,AK081116
10479971,Gm10855
10479975,ENSMUST00000083890
10482802,Cytip
10483046,Dpp4
10484203,2610301F02Rik
10484207,2610301F02Rik
10484463,Serping1
10485261,Accsl
10485784,Olfr1297
10486396,Ehd4
10487252,Gabpb1
10488912,Edem2
10489107,Samhd1
10489829,ENSMUST00000083937

10490815,Gm5841
10490872,Lrrcc1
10490989,Cp
10491091,Tnfsf10
10491846,ENSMUST00000082993
10491848,Larp1b
10492558,Smc4
10494978,Ptpn22
10495054,Rhoc
10495685,Arhgap29
10496001,Cfi
10496359,Emcn
10496872,Eltd1
10497501,Naaladl2
10497773,Mccc1
10497971,Sclt1
10498350,P2ry14
10499093,BC086805
10499189,Fcrls
10500677,Cd2
10502201,A430072C10Rik
10505270,Slc31a2
10505954,Tek
10510872,chr4:153570616:153570666:+
10514323,ENSMUSG00000073810
10514398,5830433M19Rik
10515187,Cyp4a14
10517573,Cela3b
10517664,chr4:138383291:138383391:-
10518408,Plod1 /// Myo5b
10519527,Abcb1a
10519555,Abcb1b
10519607,4930420K17Rik
10519940,ENSMUST00000082570
10520445,Dnajb6
10525439,P2rx4
10527649,6330406I15Rik
10527963,Gm10484
10528622,Asb10
10530492,Nfxl1
10530841,Igfbp7
10531288,ENSMUST00000075293 /// GENSCAN00000006129
10531407,Cxcl9
10531987,Gbp4

10534301,Gm52
10535807,Flt1
10537227,Tmem140 /// 3110062M04Rik
10538247,Npy
10540105,Tmem43
10540493,Edem1
10541318,Slc6a13
10541515,Dppa3
10544523,Rny1
10544588,Gimap3
10545217,LOC100046973
10545239,ENSMUST00000103387 /// M28833 /// AJ222611
10547740,C1s
10548401,Klrc2
10548892,Arhgdib
10548996,Slco1a4
10549377,1700034J05Rik
10550818,V1rd22
10552824,Rras
10554712,2610206C17Rik
10555323,P4ha3
10560030,Vmn2r56
10560190,Ehd2
10560242,C5ar1
10562166,Hamp2
10562192,Fxyd5
10563834,GENSCAN00000028249
10563919,ENSMUST00000101941
10563933,ENSMUST00000101908
10563943,ENSMUST00000101944
10563961,ENSMUST00000097231
10563973,ENSMUST00000101879
10563991,ENSMUST00000101803
10564057,ENSMUST00000101951
10564287,Gm7482
10565018,Iqgap1
10565994,Art2b
10566283,Olfr64
10566294,Olfr640
10566413,Olfr672
10566574,Gvin1
10566702,Olfr517
10572432,Isyna1
10573578,BC056474

10574438,Cdh5
10576049,Foxf1a
10576661,Itgb1
10578685,GENSCAN00000008377
10579894,Hhip
10580033,Cd97
10581378,Psmb10
10581388,Lcat
10582419,Pabpn1l
10583145,Tmem123
10583203,Phxr4
10584589,mmu-mir-100
10585697,Gm5121
10586172,ENSMUST00000083425
10586441,Oaz2
10588079,7420426K07Rik
10588201,ENSMUST00000083173
10588849,Amigo3
10588931,BC048562
10589535,Ngp
10590860,9030420J04Rik
10591194,Olfr850
10592449,Olfr149
10592469,Olfr985
10593050,Il10ra
10595439,ENSMUST00000083772
10595768,Pls1
10597518,Tgfbr2 /// Mib1
10597648,Myd88
10597960,Slc6a20a
10598075,NC_005089
10600169,Bgn
10600584,ENSMUST00000118391 /// GENSCAN00000019880
10601863,Gm7966
10603583,Srpx /// Rpgr
10603794,ENSMUST00000083134
10604832,mmu-mir-505
10606301,Magt1
10606389,Gm379
10606542,ENSMUST00000118182 /// GENSCAN00000047914
10607116,Ammecr1
10607848,Egfl6
10608136,ENSMUST00000101925

Appendix D. Installation steps to recreate the Galaxy development environment.

1. Download and install VirtualBox. http://www.virtualbox.org/wiki/Downloads
2. Create new virtual machine of type Linux and distribution Gentoo. Use default parameters.
3. When asked to provide installation media: retrieve iso from http://distfiles.gentoo.org/releases/x86/autobuilds/current-iso/ and save it to disk.
4. Provide VirtualBox with the path to the .iso file.
5. Follow the installation instructions in the Gentoo handbook. http://www.gentoo.org/doc/en/handbook/index.xml
6. Once the kernel is compiled, shut down the virtual machine, go to the settings for the machine in VirtualBox, and under Storage, remove the .iso from the CD drive.
7. Complete the installation instructions in the Gentoo handbook.
8. $ emerge mercurial
9. $ hg clone http://bitbucket.org/galaxy/galaxy-central
10. Shut down the virtual machine.
11. Launch the windows command line interface as an administrator.
12. cd to the VirtualBox directory.
13. VBoxManage modifyvm "VM name" --natpf1 "galaxy,tcp,127.0.0.1,8080,127.0.0.1,8080"
14. Restart the virtual machine.
15. cd to the galaxy-central directory.
16. $ sh run.sh
17. Open a browser inside Windows. Type http://127.0.0.1:8080 into the address bar.
18. Galaxy should be up and running.

Appendix E. Clustering output files.

results.cdt (from 3rd clustering run):

"GID" "*" "NAME" "GWEIGHT" "MJ10.CEL" "MJ12.CEL" "MJ22.CEL" "MJ14.CEL" "MJ11.CEL" "MJ7.CEL" "MJ9.CEL" "MJ5.CEL"
"AID" "ARRY0X" "ARRY2X" "ARRY4X" "ARRY3X" "ARRY1X" "ARRY6X" "ARRY7X" "ARRY5X"
"EWEIGHT" 1 1 1 1 1 1 1 1
"GENE15X" 10363157 10363157 1 126.925 130.282 104.795 124.493 149.596 160.89 151.181 153.327
"GENE4X" 10349174 10349174 1 295.361 293.563 294.305 286.324 364.414 387.08 347.244 334.139
"GENE8X" 10356145 10356145 1 201.235 205.045 203.436 219.234 258.308 270.831 228.776 246.238
"GENE124X" 10479975 10479975 1 323.174 345.842 332.799 377.364 421.737 439.05 416.197 380.768
"GENE162X" 10527649 10527649 1 278.336 226.024 247.783 245.06 318.618 312.781 272.196 286.3
"GENE75X" 10428509 10428509 1 275.549 266.76 236.049 344.597 413.219 326.5 371.881 321.399
"GENE86X" 10436658 10436658 1 251.163 277.533 248.14 271.896 350.491 299.3 330.295 275.43
"GENE49X" 10397428 10397428 1 341.987 319.688 315.47 333.945 422.036 393.791 377.145 382.051
"GENE115X" 10466938 10466938 1 203.285 197.05 190.768 193.93 254.25 234.423 229.169 219.002
"GENE102X" 10454828 10454828 1 239.938 223.381 229.77 251.1 328.656 275.42 283.56 275.426
"GENE163X" 10527963 10527963 1 116.815 102.592 89.9433 106.165 152.128 121.192 132.007 112.529
"GENE14X" 10363007 10363007 1 396.861 421.426 382.934 411.809 497.429 465.546 462.443 503.405
"GENE82X" 10435767 10435767 1 265.859 298.809 241.736 297.299 442.549 309.556 382.263 429.663
"GENE130X" 10487252 10487252 1 249.002 248.089 233.245 247.674 317.801 260.227 279.846 319.272
"GENE29X" 10376163 10376163 1 363.678 351.897 329.138 328.068 450.242 369.559 400.981 430.282
"GENE59X" 10408741 10408741 1 321.507 315.224 305.106 290.443 401.654 358.681 357.038 388.558
"GENE175X" 10541318 10541318 1 290.415 292.816 260.38 234.724 354.701 312.519 306.565 336.518
"GENE1X" 10346235 10346235 1 131.377 131.582 137.743

```
        140.116         190.303         161.348         150.424         189.134
"GENE144X" 10497971     10497971        1       135.101         143.438         141.954
        141.343         181.957         165.228         153.377         181.749
"GENE108X" 10461012     10461012        1       229.213         227.942         235.27 245.948
        322.075         261.48 245.077          303.853
"GENE233X" 10607116     10607116        1       159.271         146.469         156.351         167.38
        224.871         186.317         167.134         218.18
"GENE140X" 10495685     10495685        1       178.275         176.165         169.313
        173.082         239.645         199.6   190.372         210.976
"GENE133X" 10490872     10490872        1       142.93 128.325          144.304         145.695
        182.237         161.734         167.433         172.212
"GENE60X"   10408870    10408870        1       361.22 315.222          363.593         441.79
        518.289         429.003         440.192         504.94
"GENE213X" 10586441     10586441        1       271.557         269.424         269.299
        302.706         362.771         303.386         323.53 355.51
"GENE103X" 10455118     10455118        1       193.115         172.937         177.202
        196.018         244.971         244.182         210.91 244.01
"GENE173X" 10540105     10540105        1       730.411         637.893         690.006
        692.645         887.322         892.481         761.57 901.541
"GENE180X" 10548996     10548996        1       518.191         489.52 543.656          586.956
        662.04 645.392          598.526         690.85
"GENE154X" 10514398     10514398        1       146.997         129.183         145.301
        129.121         171.41 156.633          153.387         181.65
"GENE77X"   10430645    10430645        1       86.4058         104.195         95.6622
        87.8713         114.65 115.103          128.443         127.512
"GENE116X" 10467110     10467110        1       217.109         223.853         207.033         208.09
        229.059         248.122         267.475         298.299
"GENE95X"   10447097    10447097        1       125.976         142.108         129.181         106.11
        149.924         139.011         152.035         157.878
"GENE218X" 10590860     10590860        1       245.341         276.104         261.628         233.74
        319.755         284.153         283.681         335.824
"GENE123X" 10479971     10479971        1       277.198         289.836         265.665
        236.531         333.293         319.203         351.053         306.804
"GENE136X" 10491846     10491846        1       103.59 111.059          87.2698         91.5089
        120.731         124.848         125.924         110.231
"GENE3X"    10348299    10348299        1       123.291         126.043         131.389
        118.164         154.488         161.953         126.575         168.705
"GENE5X"    10349947    10349947        1       220.13 251.748          228.828         199.493
        281.674         295.005         256.695         314.853
"GENE111X" 10461723     10461723        1       114.859         123.043         129.667
        98.2324         141.442         150.426         135.883         147.076
"GENE205X" 10581378     10581378        1       505.313         530.209         558.703
        452.068         619.744         657.62 604.497          597.042
"GENE0X"    10346168    10346168        1       143.204         129.824         129.729
```

107.424 155.461 204.749 173.219 189.361
"GENE47X" 10395039 10395039 1 510.329 488.391 525.962
392.592 563.372 743.767 605.755 618.747
"GENE125X" 10483046 10483046 1 136.589 136.485 118.459 114.84
143.515 164.148 143.022 157.043
"GENE151X" 10505954 10505954 1 236.272 248.308 244.842 227.92
269.794 365.556 254.238 306.652
"GENE200X" 10576049 10576049 1 135.647 163.628 150.479
134.044 153.774 193.869 168.17 185.622
"GENE35X" 10379389 10379389 1 301.911 281.797 341.014
269.011 304.658 386.62 375.648 369.707
"GENE179X" 10548892 10548892 1 337.545 303.877 407.617
285.643 339.628 559.706 515.753 558.834
"GENE27X" 10375443 10375443 1 250.707 260.512 325.75 198.052
283.533 401.038 382.893 339.708
"GENE112X" 10463070 10463070 1 457.778 463.104 598.721 407.89
514.391 718.259 659.653 635.731
"GENE185X" 10560242 10560242 1 107.377 116.635 134.914
98.2429 120.84 156.042 143.338 146.794
"GENE11X" 10361897 10361897 1 656.019 582.806 705.448
545.704 662.17 768.337 735.883 843.448
"GENE79X" 10433101 10433101 1 268.914 238.937 281.601
204.488 280.317 327.607 302.237 341.641
"GENE58X" 10407173 10407173 1 1030.51 974.048 1171.58
941.462 1084.93 1348.31 1207.92 1478.92
"GENE87X" 10436830 10436830 1 546.533 461.972 609.932
470.619 573.89 701.333 627.38 780.365
"GENE121X" 10475199 10475199 1 962.634 953.175 1127.69
946.313 1005.19 1248.91 1192.91 1374.11
"GENE134X" 10490989 10490989 1 288.761 288.096 363.295
253.993 344.712 481.692 438.032 580.036
"GENE230X" 10606301 10606301 1 297.818 268.791 319.025
280.383 307.795 347.002 337.085 412.673
"GENE114X" 10466521 10466521 1 150.522 154.652 162.015
125.857 156.064 182.218 180.681 211.934
"GENE48X" 10396608 10396608 1 195.523 204.633 214.341
184.881 214.707 239.438 239.913 257.806
"GENE161X" 10525439 10525439 1 328.738 337.523 363.751
309.603 364.57 429.177 406.829 441.157
"GENE223X" 10597518 10597518 1 349.014 375.852 410.527
340.252 374.584 468.586 465.407 491.799
"GENE23X" 10373542 10373542 1 463.57 466.37 505.862 478.079
593.587 653.583 611.485 593.178
"GENE36X" 10379953 10379953 1 206.997 193.293 214.084

204.077 239.045 268.304 253.87 255.212
"GENE55X" 10404429 10404429 1 136.983 123.721 139.966
139.931 164.556 191.168 170.848 187.639
"GENE80X" 10433431 10433431 1 199.673 191.35 188.855 212.799
233.174 288.383 239.37 254.115
"GENE212X" 10586172 10586172 1 111.144 103.753 111.142
122.337 124.378 142.596 127.975 141.733
"GENE34X" 10379190 10379190 1 744.367 745.914 859.313
832.672 955.194 1196.16 893.454 1053.38
"GENE18X" 10366476 10366476 1 330.549 340.389 336.918
349.769 367.674 478.303 385.202 450.679
"GENE120X" 10471486 10471486 1 358.941 380.222 399.481
369.303 416.408 538.444 413.267 549.906
"GENE122X" 10477920 10477920 1 207.435 215.652 236.98 224.4 241.15
298.758 243.124 285.603
"GENE170X" 10535807 10535807 1 374.715 363.299 398.3 374.349
429.174 518.246 410.693 503.761
"GENE145X" 10498350 10498350 1 137.715 163.196 174.618
163.213 169.676 222.975 183.958 211.972
"GENE101X" 10454310 10454310 1 203.731 208.466 223.31 208.825
232.444 259.144 245.77 276.873
"GENE16X" 10364593 10364593 1 159.173 156.031 171.15 155.819
180.292 215.959 198.18 213.5
"GENE204X" 10580033 10580033 1 192.414 182.95 205.257 187.632
212.54 295.094 252.325 286.805
"GENE38X" 10384725 10384725 1 320.774 283.292 326.674
297.703 342.641 399.25 356.874 388.369
"GENE131X" 10488912 10488912 1 282.545 262.966 288.643
246.911 320.444 372.158 319.418 358.927
"GENE74X" 10426098 10426098 1 581.512 547.449 584.877
566.303 679.356 774.645 645.048 747.171
"GENE157X" 10519527 10519527 1 346.438 313.708 341.228
331.526 380.636 485.804 391.565 453.205
"GENE206X" 10581388 10581388 1 573.068 565.518 602.143
567.934 656.703 775.606 666.202 751.09
"GENE126X" 10484203 10484203 1 275.751 237.035 302.438
262.091 343.253 415.924 323.699 443.479
"GENE66X" 10416657 10416657 1 246.872 258.468 292.281
247.559 276.759 375.795 297.098 348.279
"GENE61X" 10412298 10412298 1 178.588 177.065 206.89 166.363
211.73 275.295 228.78 271.82
"GENE88X" 10436841 10436841 1 311.675 307.957 361.602
275.539 343.322 446.402 371.829 427.162
"GENE224X" 10597648 10597648 1 266.642 235.235 298.781

231.692 289.95 380.407 307.057 360.405
"GENE105X" 10458314 10458314 1 305.118 287.758 378.191
275.104 325.651 491.601 406.071 431.75
"GENE159X" 10519607 10519607 1 746.706 706.299 899.291
745.533 803.353 1068.78 956.481 1020.58
"GENE182X" 10552824 10552824 1 970.153 881.897 1075.48
975.852 1017.72 1359.73 1159.82 1344.02
"GENE208X" 10583145 10583145 1 408.438 346.541 517.973
395.918 445.836 740.551 567.39 716.299
"GENE129X" 10486396 10486396 1 255.589 237.352 285.15 230.623
310.4 331.066 283.494 312.873
"GENE158X" 10519555 10519555 1 92.9394 95.3617 128.709
96.8739 129.173 148.698 116.647 136.008
"GENE183X" 10555323 10555323 1 229.473 195.042 237.784
209.322 249.459 285.494 244.156 261.713
"GENE201X" 10576661 10576661 1 433.543 367.526 451.814
404.802 463.567 515.157 503.711 531.429
"GENE210X" 10584589 10584589 1 96.2586 83.2235 106.942
94.6278 119.292 118.187 114.036 126.545
"GENE57X" 10405785 10405785 1 165.917 156.57 137.967 154.696
177.14 175.614 172.121 205.881
"GENE127X" 10484207 10484207 1 286.566 258.81 239.989 263.874
308.031 342.669 302.846 402.345
"GENE98X" 10452404 10452404 1 251.208 221.319 232.786
252.228 286.905 280.275 248.459 329.986
"GENE143X" 10497773 10497773 1 365.329 304.505 324.973
352.166 409.831 394.115 362.473 478.691
"GENE62X" 10413803 10413803 1 232.321 222.772 232.42 237.687
282.146 255.431 265.19 324.068
"GENE9X" 10358339 10358339 1 226.618 207.847 239.399
228.587 290.841 287.291 275.181 418.037
"GENE2X" 10346551 10346551 1 144.611 138.737 149.231
147.274 165.752 166.894 168.881 217.809
"GENE17X" 10365974 10365974 1 446.71 393.832 443.249 451.516
544.525 547.182 563.084 886.046
"GENE141X" 10496359 10496359 1 314.886 325.186 334.884
347.342 412.09 408.824 397.539 671.535
"GENE118X" 10469255 10469255 1 212.532 188.579 223.731
209.656 232.521 255.094 271.531 375.281
"GENE42X" 10388430 10388430 1 310.268 314.81 346.079 311.295
367.612 368.161 371.779 467.406
"GENE139X" 10495054 10495054 1 258.047 242.304 294.027 236.85
298.866 308.167 295.924 390.892
"GENE100X" 10454192 10454192 1 163.846 174.708 279.483

196.894 1888.13 1260.79 239.731 11677.7
"GENE46X" 10392440 10392440 1 196.162 188.338 204.98 170.416
202.58 256.648 224.772 301.459
"GENE99X" 10453057 10453057 1 130.576 134.483 146.935 127.91
148.313 162.995 142.386 185.603
"GENE199X" 10574438 10574438 1 200.48 199.631 213.18 186.217 230.91
268.527 225.612 297.556
"GENE166X" 10530841 10530841 1 817.394 820.644 980.716
858.102 1032.13 1531.92 1025.64 1955.86
"GENE84X" 10436500 10436500 1 248.692 244.794 251.35 251.188
280.666 309.063 265.085 384.99
"GENE109X" 10461115 10461115 1 383.899 397.416 387.524
409.709 472.027 553.643 408.95 701.315
"GENE195X" 10565018 10565018 1 287.525 287.968 298.026
289.101 353.289 362.298 320.731 442.834
"GENE225X" 10597960 10597960 1 635.663 587.647 600.264
563.642 684.116 741.209 661.58 988.22
"GENE56X" 10405587 10405587 1 289.057 282.023 332.799
311.629 319.037 423.125 362.457 478.8
"GENE156X" 10518408 10518408 1 423.316 369.204 449.699
409.741 446.584 545.03 473.557 579.602
"GENE94X" 10445909 10445909 1 211.549 186.938 217.2 213.675
223.443 263.72 235.208 293.743
"GENE227X" 10600169 10600169 1 403.532 344.426 398.932
417.223 428.644 552.009 458.412 612.609
"GENE209X" 10583203 10583203 1 175.368 130.677 165.177
142.287 175.236 260.665 171.461 285.097
"GENE43X" 10389025 10389025 1 371.866 296.646 352.234
345.511 354.326 413.949 413.388 477.202
"GENE85X" 10436590 10436590 1 129.681 105.205 120.559
126.437 134.481 138.76 137.865 172.301
"GENE81X" 10435266 10435266 1 195.81 167.678 173.819 178.049
176.23 249.27 231.12 252.424
"GENE19X" 10368289 10368289 1 313.618 271.142 265.872
312.437 327.864 387.951 380.913 450.915
"GENE222X" 10595768 10595768 1 271.523 223.739 237.387
255.502 269.674 301.229 291.6 317.645
"GENE229X" 10604832 10604832 1 372.848 320.897 341.415
359.926 396.809 474.78 445.651 484.208
"GENE45X" 10392221 10392221 1 137.804 159.919 173.327 164.93
187.987 208.035 165.234 204.954
"GENE198X" 10573578 10573578 1 1322.56 1462 1658.33 1727.72
1863.98 2053.82 1768.77 2025.6
"GENE90X" 10438517 10438517 1 866.433 827.876 880.22 1022.08

1022.5 1118.45 954.383 1240.22
"GENE117X" 10469151 10469151 1 193.689 227.962 225.454
230.939 251.574 292.236 233.182 329.539
"GENE142X" 10496872 10496872 1 315.731 329.544 350.383
374.306 398.723 519.255 397.183 539.056
"GENE197X" 10572432 10572432 1 469.623 494.878 525.62 539.015
578.047 640.495 541.474 667.876
"GENE119X" 10469358 10469358 1 104.47 115.282 108.749 127.297
137.124 141.641 139.047 147.357
"GENE172X" 10538247 10538247 1 1925.64 1924.04 1950.01
2118.41 2404.98 2340.73 2405.3 2493.73
"GENE184X" 10560190 10560190 1 150.023 167.691 157.841
170.865 183.254 199.701 192.57 217.11
"GENE39X" 10385504 10385504 1 93.2089 78.589 99.5812 86.4375
98.3774 174.028 115.998 122.986
"GENE174X" 10540493 10540493 1 435.336 392.019 541.862
431.147 506.349 737.485 599.709 587.427
"GENE28X" 10376060 10376060 1 199.087 185.19 199.402 135.635
199.281 352.922 299.142 239.498
"GENE31X" 10376326 10376326 1 453.879 409.914 584.814
225.077 489.345 1306.4 1095.09 692.553
"GENE30X" 10376324 10376324 1 322.653 230.835 331.254
169.327 335.92 894.594 683.321 398.561
"GENE135X" 10491091 10491091 1 182.086 171.173 188.225
150.202 199.752 347.115 274.316 230.031
"GENE69X" 10420308 10420308 1 90.0385 90.6749 87.5616 76.749
98.6265 165.18 122.213 99.2988
"GENE168X" 10531987 10531987 1 228.448 197.744 292.458
117.956 272.155 1213.04 719.68 381.16
"GENE196X" 10566574 10566574 1 437.024 420.586 442.203
319.685 470.372 622.387 560.606 459.98
"GENE10X" 10360391 10360391 1 178.889 132.011 195.574
97.5048 155.778 400.782 335.259 275.885
"GENE177X" 10544588 10544588 1 99.8529 77.2327 93.1162
76.0402 97.4761 157.884 134.845 125.402
"GENE128X" 10484463 10484463 1 407.025 301.911 558.162
235.059 420.036 1149.13 732.533 783.357
"GENE132X" 10489107 10489107 1 521.547 441.006 565.929
355.853 522.12 1076.3 763.866 720.598
"GENE40X" 10385511 10385511 1 271.469 249.567 298.292
182.223 281.494 433.876 370.951 352.789
"GENE220X" 10593050 10593050 1 258.809 250.816 296.629
211.542 268.779 398.168 358.515 318.284
"GENE150X" 10505270 10505270 1 966.555 933.435 1051.84

895.866 1027.44 1270.25 1185.6 1154.35
"GENE91X" 10444229 10444229 1 385.993 306.485 396.941
267.995 423.318 550.269 608.925 415.679
"GENE93X" 10444298 10444298 1 377.412 343.67 491.954 323.604
479.366 1048.12 1267.53 595.417
"GENE96X" 10450154 10450154 1 395.292 336.139 791.621
179.852 714.52 2079.73 2523.68 898.918
"GENE167X" 10531407 10531407 1 448.799 419.58 592.869 234.848
547.999 1183.96 1385.67 621.218
"GENE92X" 10444291 10444291 1 354.286 372.508 830.15 171.042
712.913 2377.2 2364.93 935.065
"GENE104X" 10456005 10456005 1 467.393 430.146 1092.54
154.601 837.14 3070.08 2978.31 1134.62
"GENE187X" 10562192 10562192 1 1029.86 1092.08 1155.7 981.791
1178.53 1465.09 1476.01 1188.58
"GENE138X" 10494978 10494978 1 89.3625 91.7694 102.253
82.8552 98.1526 180.229 162.272 116.825
"GENE171X" 10537227 10537227 1 140.723 159.982 157.785
133.313 170.421 296.714 275.906 182.038
"GENE148X" 10500677 10500677 1 109.92 110.442 111.741 117.622
113.34 155.496 141.025 138.721
"GENE54X" 10404389 10404389 1 110.042 102.005 107.773
92.8874 141.063 186.03 144.084 105.059
"GENE178X" 10545239 10545239 1 80.1031 82.9875 204.477
69.4619 5108.42 7431.19 7322.11 127.608
"GENE165X" 10530492 10530492 1 110.625 123.39 150.14 131.365
148.867 162.081 176.012 154.682
"GENE64X" 10415952 10415952 1 120.904 94.945 116.799 111.86
116.116 136.345 155.451 146.619
"GENE137X" 10492558 10492558 1 121.59 97.4129 111.404 101.983
114.954 140.04 140.897 140.514
"GENE203X" 10579894 10579894 1 214.671 182.051 187.761
155.925 201.245 215.589 279.165 251.292
"GENE41X" 10387219 10387219 1 194.34 215.089 189.027 218.302
134.802 171.828 185.557 180.462
"GENE113X" 10466288 10466288 1 137.246 168.636 145.561
163.549 110.377 131.74 136.234 132.733
"GENE24X" 10373606 10373606 1 99.0662 116.621 128.506
138.144 92.0512 91.935 106.785 95.2682
"GENE50X" 10399581 10399581 1 175.98 168.932 197.589 213.108
154.697 135.089 168.981 140.996
"GENE160X" 10520445 10520445 1 155.444 178.668 187.204
182.947 127.083 112.535 159.839 154.005
"GENE26X" 10375129 10375129 1 183.077 199.678 193.734

195.391 148.75 162.635 184.765 151.862
"GENE7X" 10354414 10354414 1 160.803 171.729 166.094
164.015 122.299 146.927 150.024 128.629
"GENE68X" 10419900 10419900 1 121.111 130.797 130.998
129.579 98.8009 112.733 116.924 95.9675
"GENE32X" 10376532 10376532 1 142.241 150.924 152.344
161.413 115.008 132.551 136.403 116.309
"GENE231X" 10606389 10606389 1 155.812 163.35 162.493 177.813
117.894 150.632 149.604 128.076
"GENE72X" 10424287 10424287 1 121.636 140.234 138.897 134.05
101.028 107.181 118.569 109.369
"GENE221X" 10595439 10595439 1 277.518 317.51 313.538 310.727
254.99 257.715 277.798 233.453
"GENE12X" 10362450 10362450 1 173.602 184.13 169.604 152.439
128.172 156.091 154.435 121.51
"GENE152X" 10510872 10510872 1 357.4 381.93 365.573 334.157
262.618 324.647 334.214 282.331
"GENE83X" 10436200 10436200 1 289.729 335.713 326.668
288.095 235.584 271.477 288.479 237.677
"GENE44X" 10390931 10390931 1 1309.67 1405.75 1518.72
1335.98 947.737 1157.38 1316.07 957.572
"GENE63X" 10414706 10414706 1 573.625 600.107 662.516
570.959 406.505 524.019 576.779 449.549
"GENE71X" 10423654 10423654 1 143.301 147.615 154.895
133.766 108.344 126.047 135.456 114.782
"GENE78X" 10430899 10430899 1 292.215 291.556 319.997
275.672 215.176 273.828 270.821 235.089
"GENE153X" 10514323 10514323 1 245.942 258.586 271.418
229.329 182.408 237.066 219.783 185.185
"GENE53X" 10403246 10403246 1 460.343 469.823 493.779 434.41
361.391 398.947 417.495 344.646
"GENE202X" 10578685 10578685 1 204.851 199.34 226.791 188.711
163.125 185.114 187.203 156.554
"GENE25X" 10373610 10373610 1 688.659 802.236 883.64 699.577
498.912 622.947 636.328 495.252
"GENE232X" 10606542 10606542 1 2698.22 2924.01 3298.67
2928.85 2134.4 2473.29 2466.02 2265.6
"GENE52X" 10401136 10401136 1 4624.84 4248.59 6980.64
1568.15 326.816 388.823 333.411 388.316
"GENE51X" 10401109 10401109 1 217.756 221.478 238.712
193.745 187.435 186.467 196.663 166.189
"GENE207X" 10582419 10582419 1 196.246 188.486 206.82 169.343
153.727 156.997 158.274 140.789
"GENE65X" 10416271 10416271 1 542.269 527.517 557.812

563.039 413.373 481.276 483.576 445.244
"GENE67X" 10418718 10418718 1 165.69 158.115 172.116 167.924
121.111 128.718 136.937 131.986
"GENE13X" 10362672 10362672 1 2449.39 2273.3 2517.97 2248.99
1850.46 2080.6 2026.13 1873.01
"GENE106X" 10458882 10458882 1 280.88 258.2 300.895 264.712
209.272 230.71 229.884 229.503
"GENE194X" 10564057 10564057 1 1903.11 1728.13 2226.95
1764.41 1069.94 1383.66 1419.58 1255.19
"GENE188X" 10563919 10563919 1 2360.19 1873.34 2148.02
2049.29 1293.98 1690.41 1734.61 1564.08
"GENE193X" 10563991 10563991 1 1562.6 1241.51 1436.97 1290.22
859.797 1114.12 1136.85 947.61
"GENE190X" 10563943 10563943 1 3480.32 3044.16 3363.88
3149.41 2328.01 2613.38 2785.47 2551.78
"GENE191X" 10563961 10563961 1 834.932 627.744 787.213
693.399 389.086 528.523 509.662 461.579
"GENE192X" 10563973 10563973 1 2438.16 2153.8 2212.04 2156.23
1434.42 1897.42 1790.81 1692.04
"GENE21X" 10370327 10370327 1 258.475 258.527 253.397
245.316 202.523 222.394 216.745 166.415
"GENE22X" 10373358 10373358 1 104.747 125.167 112.72 117.351
94.4384 105.652 89.8438 79.7461
"GENE164X" 10528622 10528622 1 167.415 190.481 181.687
187.312 147.884 160.697 153.129 132.755
"GENE107X" 10459618 10459618 1 388.522 433.039 348.449
385.794 313.73 315.075 313.891 287.767
"GENE215X" 10588201 10588201 1 194.057 227.287 187.025
211.388 175.013 167.76 163.269 139.011
"GENE37X" 10381334 10381334 1 131.167 127.721 115.049
112.796 85.0805 101.394 104.381 102.547
"GENE73X" 10424695 10424695 1 285.91 308.73 264.38 254.044 205.244
226.069 243.658 225.839
"GENE214X" 10588079 10588079 1 251.873 276.943 240.637
247.294 186.277 226.075 218.995 210.704
"GENE228X" 10600584 10600584 1 121.755 151.064 118.188
121.797 100.162 114.272 100.587 112.65
"GENE216X" 10588849 10588849 1 201.202 224.3 210.34 202.313
179.309 164.613 178.59 183.184
"GENE217X" 10588931 10588931 1 136.31 154.745 148.66 133.382
118.261 121.703 114.552 117.054
"GENE97X" 10450191 10450191 1 192.674 188.424 217.962
223.858 156.632 176.607 142.508 176.371
"GENE155X" 10517573 10517573 1 173.996 176.678 179.214

214.866 141.25 164.336 151.827 152.189
"GENE70X" 10420899 10420899 1 151.237 147.56 173.09 157.975
127.649 138.942 118.905 120.099
"GENE219X" 10592449 10592449 1 180.545 199.345 209.707
202.826 158.698 191.756 158.365 156.549
"GENE33X" 10376885 10376885 1 2260.67 1654.46 1777.8 1683.42
1714.11 1186.29 1266.79 1371.9
"GENE110X" 10461156 10461156 1 494.588 385.606 412.935
416.464 377.997 332.215 301.273 350.09
"GENE76X" 10429638 10429638 1 242.876 242.065 265.689
210.781 207.06 214.147 171.643 143.18
"GENE6X" 10351041 10351041 1 249.956 217.339 262.167
235.445 220.089 213.371 179.74 185.318
"GENE89X" 10436890 10436890 1 442.848 368.241 471.328
374.795 354.546 335.901 271.236 302.775
"GENE149X" 10502201 10502201 1 136.674 128.079 143.471
137.514 127.055 123.279 108.326 100.047
"GENE169X" 10534301 10534301 1 221.18 218.489 217.785 208.864
189.169 181.071 167.956 171.471
"GENE181X" 10549377 10549377 1 127.401 117.994 118.001
118.577 105.072 103.919 89.937 94.8472
"GENE146X" 10499093 10499093 1 152.698 154.626 145.008
151.572 128.89 102.408 126.415 119.107
"GENE186X" 10562166 10562166 1 316.253 291.962 240.791
308.143 212.924 174.335 254.713 211.244
"GENE20X" 10368370 10368370 1 183.858 159.899 149.272
160.164 134.512 136.988 146.446 120.926
"GENE189X" 10563933 10563933 1 447.709 322.895 371.59 381.392
271.535 309.944 331.509 286.977
"GENE176X" 10544523 10544523 1 1357.66 1135.14 1278.35
1369.86 983.696 750.957 1008.72 767.407
"GENE211X" 10585697 10585697 1 371.658 315.41 339.121 381.489
300.168 285.993 306.417 263.737
"GENE226X" 10598075 10598075 1 1403.53 971.376 1140.45
1357.73 913.399 746.95 834.938 820.567
"GENE234X" 10608136 10608136 1 370.655 339.137 388.852
370.451 307.672 297.651 338.903 273.628
"GENE147X" 10499189 10499189 1 182.598 203.063 273.147 197.32
187.837 158.841 134.777 167.343

results.gtr (from 3rd clustering run):

"NODE1X" "GENE104X" "GENE92X" 0.9995607771
"NODE2X" "GENE96X" "GENE93X" 0.9982026976

"NODE3X" "GENE208X" "GENE182X" 0.9981043183
"NODE4X" "GENE17X" "GENE2X" 0.9976510938
"NODE5X" "GENE227X" "GENE94X" 0.9959476102
"NODE6X" "GENE167X" "NODE2X" 0.9955724546
"NODE7X" "GENE9X" "NODE4X" 0.9954320925
"NODE8X" "GENE143X" "GENE98X" 0.9952659377
"NODE9X" "GENE109X" "GENE84X" 0.9946171978
"NODE10X" "GENE31X" "GENE28X" 0.994571372
"NODE11X" "GENE171X" "GENE138X" 0.9944712497
"NODE12X" "GENE194X" "GENE106X" 0.9938813404
"NODE13X" "GENE204X" "GENE16X" 0.9934006598
"NODE14X" "GENE87X" "GENE58X" 0.9933537571
"NODE15X" "GENE132X" "GENE128X" 0.9932346796
"NODE16X" "GENE135X" "GENE30X" 0.9929019826
"NODE17X" "GENE168X" "GENE69X" 0.9926379585
"NODE18X" "GENE193X" "GENE188X" 0.9925103215
"NODE19X" "GENE108X" "GENE233X" 0.9922999145
"NODE20X" "GENE63X" "GENE44X" 0.992082349
"NODE21X" "GENE224X" "GENE88X" 0.9917679554
"NODE22X" "GENE161X" "GENE48X" 0.9917537713
"NODE23X" "GENE187X" "NODE1X" 0.9917537008
"NODE24X" "GENE191X" "GENE190X" 0.9917278834
"NODE25X" "GENE220X" "GENE40X" 0.9916395795
"NODE26X" "GENE112X" "GENE27X" 0.9915940328
"NODE27X" "GENE144X" "GENE1X" 0.9912741481
"NODE28X" "GENE141X" "NODE7X" 0.9911974603
"NODE29X" "GENE206X" "GENE157X" 0.9911367164
"NODE30X" "NODE16X" "NODE10X" 0.990601087
"NODE31X" "GENE159X" "GENE105X" 0.9903936847
"NODE32X" "GENE134X" "GENE121X" 0.9899806579
"NODE33X" "NODE24X" "NODE18X" 0.9897074035
"NODE34X" "GENE177X" "GENE10X" 0.9896688939
"NODE35X" "NODE21X" "GENE61X" 0.9895746041
"NODE36X" "GENE173X" "GENE103X" 0.9895593068
"NODE37X" "NODE29X" "GENE74X" 0.9894897418
"NODE38X" "GENE55X" "GENE36X" 0.9884708496
"NODE39X" "NODE11X" "NODE23X" 0.9883848051
"NODE40X" "GENE150X" "NODE25X" 0.9883703871
"NODE41X" "GENE139X" "GENE42X" 0.9880786504
"NODE42X" "GENE156X" "NODE5X" 0.9880682521
"NODE43X" "GENE142X" "GENE197X" 0.9879409679
"NODE44X" "GENE170X" "GENE122X" 0.9877173712
"NODE45X" "GENE79X" "GENE11X" 0.9876837512
"NODE46X" "GENE185X" "NODE26X" 0.9873198096

"NODE47X"	"GENE32X"	"GENE231X"	0.9870835202
"NODE48X"	"GENE199X"	"GENE99X"	0.986714133
"NODE49X"	"GENE68X"	"GENE7X"	0.9862939286
"NODE50X"	"GENE195X"	"NODE9X"	0.985957778
"NODE51X"	"GENE179X"	"GENE35X"	0.9859232325
"NODE52X"	"GENE130X"	"GENE82X"	0.9856998256
"NODE53X"	"NODE44X"	"GENE120X"	0.9854890213
"NODE54X"	"GENE115X"	"GENE49X"	0.984991883
"NODE55X"	"NODE42X"	"GENE56X"	0.9848774324
"NODE56X"	"GENE89X"	"GENE6X"	0.9847758586
"NODE57X"	"GENE131X"	"GENE38X"	0.9847065977
"NODE58X"	"NODE32X"	"NODE14X"	0.9846469873
"NODE59X"	"GENE223X"	"NODE22X"	0.9846018531
"NODE60X"	"GENE118X"	"NODE28X"	0.9844395245
"NODE61X"	"GENE71X"	"NODE20X"	0.9841388742
"NODE62X"	"GENE229X"	"GENE222X"	0.9836093722
"NODE63X"	"GENE166X"	"NODE48X"	0.9835052387
"NODE64X"	"GENE153X"	"GENE78X"	0.9834192746
"NODE65X"	"GENE164X"	"GENE22X"	0.9825638118
"NODE66X"	"NODE6X"	"NODE39X"	0.9823536688
"NODE67X"	"NODE12X"	"GENE13X"	0.9822207418
"NODE68X"	"GENE66X"	"NODE35X"	0.9821949123
"NODE69X"	"GENE207X"	"GENE51X"	0.98209097
"NODE70X"	"GENE181X"	"GENE169X"	0.9819911821
"NODE71X"	"NODE50X"	"GENE225X"	0.9816765268
"NODE72X"	"GENE111X"	"GENE205X"	0.9815784946
"NODE73X"	"NODE15X"	"NODE34X"	0.9812520061
"NODE74X"	"NODE3X"	"NODE31X"	0.9808569016
"NODE75X"	"GENE202X"	"GENE53X"	0.980816528
"NODE76X"	"NODE40X"	"NODE73X"	0.9805054155
"NODE77X"	"NODE37X"	"NODE57X"	0.9803924239
"NODE78X"	"NODE17X"	"NODE30X"	0.9801642884
"NODE79X"	"GENE117X"	"NODE43X"	0.979683267
"NODE80X"	"GENE192X"	"NODE33X"	0.9796340943
"NODE81X"	"NODE63X"	"GENE46X"	0.9794571287
"NODE82X"	"GENE230X"	"NODE58X"	0.9794266054
"NODE83X"	"NODE19X"	"NODE27X"	0.9788403836
"NODE84X"	"GENE232X"	"GENE25X"	0.9786596307
"NODE85X"	"GENE152X"	"GENE12X"	0.978287253
"NODE86X"	"GENE59X"	"GENE29X"	0.9782586613
"NODE87X"	"NODE53X"	"GENE18X"	0.9781816063
"NODE88X"	"GENE26X"	"NODE49X"	0.9779478132
"NODE89X"	"GENE101X"	"NODE13X"	0.9779459913
"NODE90X"	"GENE127X"	"GENE57X"	0.9778299586

"NODE91X"	"NODE62X"	"GENE19X"	0.9777048481
"NODE92X"	"GENE172X"	"GENE119X"	0.9774686662
"NODE93X"	"NODE74X"	"NODE68X"	0.9763705312
"NODE94X"	"GENE47X"	"GENE0X"	0.9760406858
"NODE95X"	"GENE213X"	"GENE60X"	0.9757774131
"NODE96X"	"GENE126X"	"NODE77X"	0.9757648105
"NODE97X"	"NODE38X"	"GENE23X"	0.9756764557
"NODE98X"	"GENE107X"	"GENE215X"	0.9745012545
"NODE99X"	"GENE110X"	"GENE33X"	0.9743172063
"NODE100X"	"GENE67X"	"GENE65X"	0.9742390006
"NODE101X"	"NODE41X"	"NODE60X"	0.973565395
"NODE102X"	"GENE174X"	"GENE39X"	0.9733149898
"NODE103X"	"GENE176X"	"GENE211X"	0.9723576403
"NODE104X"	"GENE73X"	"GENE37X"	0.9722429015
"NODE105X"	"GENE209X"	"NODE55X"	0.9721911082
"NODE106X"	"GENE113X"	"GENE41X"	0.9721520372
"NODE107X"	"GENE114X"	"NODE59X"	0.9714378343
"NODE108X"	"GENE129X"	"GENE158X"	0.9713497162
"NODE109X"	"NODE51X"	"NODE46X"	0.9711654397
"NODE110X"	"GENE34X"	"NODE87X"	0.969746161
"NODE111X"	"GENE85X"	"GENE43X"	0.9696565665
"NODE112X"	"NODE64X"	"NODE75X"	0.9691256373
"NODE113X"	"GENE137X"	"GENE64X"	0.9687477963
"NODE114X"	"GENE102X"	"NODE54X"	0.9686646329
"NODE115X"	"NODE112X"	"NODE61X"	0.9682106832
"NODE116X"	"NODE96X"	"NODE89X"	0.9678345127
"NODE117X"	"GENE198X"	"GENE45X"	0.9678123465
"NODE118X"	"GENE196X"	"NODE78X"	0.9674591221
"NODE119X"	"GENE80X"	"GENE212X"	0.9670726429
"NODE120X"	"NODE81X"	"NODE71X"	0.9670351264
"NODE121X"	"GENE175X"	"NODE86X"	0.9669515499
"NODE122X"	"GENE52X"	"NODE69X"	0.9656997945
"NODE123X"	"NODE82X"	"NODE45X"	0.9655507235
"NODE124X"	"GENE83X"	"NODE85X"	0.9646358354
"NODE125X"	"NODE66X"	"GENE91X"	0.9642143352
"NODE126X"	"GENE201X"	"GENE210X"	0.9640479334
"NODE127X"	"GENE226X"	"NODE103X"	0.9640022961
"NODE128X"	"NODE47X"	"NODE88X"	0.9638814897
"NODE129X"	"GENE62X"	"NODE101X"	0.9631918892
"NODE130X"	"GENE145X"	"NODE110X"	0.9631580676
"NODE131X"	"NODE118X"	"NODE76X"	0.9629909233
"NODE132X"	"GENE214X"	"NODE104X"	0.9614773132
"NODE133X"	"GENE90X"	"NODE79X"	0.961017729
"NODE134X"	"GENE221X"	"GENE72X"	0.9608229189

"NODE135X" "NODE52X" "GENE14X" 0.9607772125
"NODE136X" "NODE80X" "NODE67X" 0.9605279969
"NODE137X" "NODE107X" "NODE123X" 0.9603096765
"NODE138X" "GENE81X" "NODE91X" 0.9602669256
"NODE139X" "NODE93X" "NODE116X" 0.9601322565
"NODE140X" "GENE184X" "NODE92X" 0.9585878407
"NODE141X" "GENE140X" "NODE83X" 0.9573496346
"NODE142X" "NODE115X" "NODE84X" 0.957205889
"NODE143X" "NODE36X" "GENE180X" 0.9567892653
"NODE144X" "NODE108X" "GENE183X" 0.9552448939
"NODE145X" "GENE5X" "GENE3X" 0.95521227
"NODE146X" "GENE133X" "NODE95X" 0.9547140772
"NODE147X" "GENE8X" "GENE4X" 0.9545003171
"NODE148X" "NODE105X" "NODE120X" 0.9540248958
"NODE149X" "NODE100X" "NODE136X" 0.9540239648
"NODE150X" "NODE130X" "NODE139X" 0.9529438948
"NODE151X" "GENE178X" "GENE54X" 0.9516167141
"NODE152X" "NODE102X" "NODE131X" 0.9512940944
"NODE153X" "GENE149X" "NODE56X" 0.9511342778
"NODE154X" "GENE123X" "GENE136X" 0.9511281068
"NODE155X" "NODE90X" "NODE8X" 0.9504247374
"NODE156X" "NODE119X" "NODE97X" 0.9492472353
"NODE157X" "GENE100X" "NODE129X" 0.9486658443
"NODE158X" "GENE70X" "GENE219X" 0.9482244632
"NODE159X" "GENE217X" "GENE216X" 0.9480446854
"NODE160X" "GENE125X" "NODE94X" 0.9473763476
"NODE161X" "GENE50X" "GENE24X" 0.9459509873
"NODE162X" "GENE186X" "GENE146X" 0.9439340194
"NODE163X" "NODE134X" "NODE128X" 0.943830031
"NODE164X" "NODE133X" "NODE117X" 0.9431117483
"NODE165X" "NODE142X" "NODE124X" 0.9430486327
"NODE166X" "NODE111X" "NODE138X" 0.9427484144
"NODE167X" "NODE114X" "GENE163X" 0.9416952391
"NODE168X" "GENE124X" "NODE147X" 0.9381687594
"NODE169X" "NODE98X" "NODE65X" 0.9373148027
"NODE170X" "GENE218X" "GENE95X" 0.9372236891
"NODE171X" "NODE121X" "NODE135X" 0.9360900557
"NODE172X" "NODE125X" "NODE152X" 0.935017333
"NODE173X" "NODE148X" "NODE157X" 0.9344892061
"NODE174X" "GENE76X" "NODE153X" 0.9327469617
"NODE175X" "NODE146X" "NODE141X" 0.9324061131
"NODE176X" "NODE156X" "NODE150X" 0.9323378212
"NODE177X" "GENE189X" "GENE20X" 0.9320846202
"NODE178X" "GENE75X" "GENE86X" 0.9305624326

"NODE179X" "NODE109X" "NODE137X" 0.9292778818
"NODE180X" "GENE151X" "GENE200X" 0.9272719452
"NODE181X" "GENE116X" "GENE77X" 0.9267557419
"NODE182X" "GENE203X" "NODE113X" 0.9265234728
"NODE183X" "NODE144X" "NODE126X" 0.92576809
"NODE184X" "NODE143X" "NODE175X" 0.924872206
"NODE185X" "NODE169X" "GENE21X" 0.9231703454
"NODE186X" "NODE70X" "NODE174X" 0.9230505778
"NODE187X" "NODE127X" "GENE234X" 0.9227303575
"NODE188X" "GENE155X" "GENE97X" 0.9214750448
"NODE189X" "NODE180X" "NODE160X" 0.921165115
"NODE190X" "NODE155X" "NODE173X" 0.9199425608
"NODE191X" "GENE148X" "NODE172X" 0.9196312187
"NODE192X" "NODE165X" "NODE163X" 0.9174502484
"NODE193X" "NODE183X" "NODE176X" 0.9173741158
"NODE194X" "NODE167X" "NODE178X" 0.9145219038
"NODE195X" "NODE193X" "NODE179X" 0.9096780379
"NODE196X" "NODE72X" "NODE145X" 0.9087853347
"NODE197X" "GENE15X" "NODE168X" 0.9081939236
"NODE198X" "GENE154X" "NODE184X" 0.9075535569
"NODE199X" "NODE164X" "NODE140X" 0.9061022285
"NODE200X" "NODE171X" "NODE198X" 0.9035562154
"NODE201X" "NODE166X" "NODE190X" 0.9022551963
"NODE202X" "NODE161X" "GENE160X" 0.9005245488
"NODE203X" "NODE132X" "GENE228X" 0.9000089256
"NODE204X" "NODE196X" "NODE189X" 0.8981995445
"NODE205X" "NODE187X" "NODE177X" 0.8970194199
"NODE206X" "NODE122X" "NODE149X" 0.8936208692
"NODE207X" "NODE188X" "NODE158X" 0.8905793315
"NODE208X" "NODE206X" "NODE192X" 0.8896702015
"NODE209X" "NODE181X" "NODE170X" 0.8873523144
"NODE210X" "NODE195X" "NODE204X" 0.8858021925
"NODE211X" "NODE203X" "NODE159X" 0.8851319435
"NODE212X" "NODE151X" "NODE191X" 0.882833612
"NODE213X" "NODE199X" "NODE201X" 0.8779649488
"NODE214X" "NODE205X" "NODE162X" 0.8741031957
"NODE215X" "GENE162X" "NODE197X" 0.8621236799
"NODE216X" "NODE99X" "NODE186X" 0.8612534087
"NODE217X" "NODE185X" "NODE211X" 0.8595187482
"NODE218X" "NODE215X" "NODE194X" 0.8585669842
"NODE219X" "NODE213X" "NODE210X" 0.8583517791
"NODE220X" "NODE212X" "GENE165X" 0.8465234636
"NODE221X" "NODE208X" "NODE217X" 0.8409693102
"NODE222X" "NODE218X" "NODE200X" 0.8399584883

"NODE223X" "NODE220X" "NODE182X" 0.8266606106
"NODE224X" "NODE154X" "NODE209X" 0.823162754
"NODE225X" "NODE202X" "NODE106X" 0.8195090871
"NODE226X" "NODE207X" "NODE221X" 0.8142581692
"NODE227X" "NODE216X" "NODE214X" 0.8130842934
"NODE228X" "NODE225X" "NODE226X" 0.795222379
"NODE229X" "NODE224X" "NODE222X" 0.7831072441
"NODE230X" "NODE223X" "NODE219X" 0.7824610125
"NODE231X" "NODE228X" "NODE227X" 0.7754305726
"NODE232X" "NODE229X" "NODE230X" 0.7328698664
"NODE233X" "GENE147X" "NODE231X" 0.7196844656
"NODE234X" "NODE233X" "NODE232X" 0

Appendix F: Treeview Documentation

All content taken from Java Treeview User's Manual with minor reformatting,
http://jtreeview.sourceforge.net/docs/JTVUserManual/single.html

Notes added by author are in negative.

Url Settings

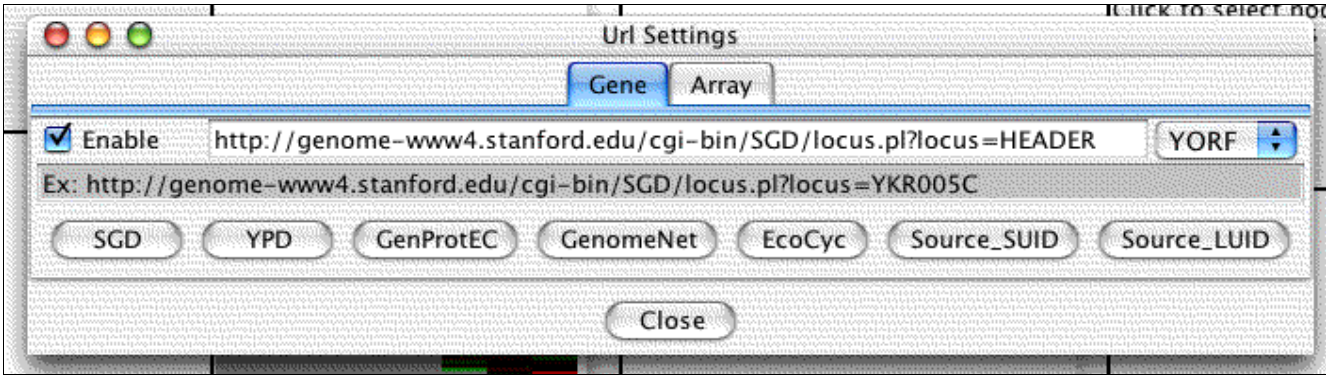

Figure F.1. Url Settings Dialog

Url Settings allow you to select from one of the presets, or to directly edit the url string. There is a special substring of the url string, "HEADER", which is replaced by a partcular gene or url header which you select from the pulldown. The default is to either use the first column for a pcl, or the second column for a cdt. In the original Eisen layout, this is the YORF column.

There is also a checkbox which allows you to disable url linking entirely. Whether this box is checked initially is determined by the default url presets.

What exactly the url settings are used for depends on the view. Generally, clicking on a gene will cause the url for that gene to be loaded in an external browser.

This feature is untested in our applet.

Dendrogram Pixel Settings

This fairly complicated dialog has three major parts. The first part allows you to set the pixel scaling for the global and zoom views. The second part allows you to set the contrast. The third part allows you to set the color settings for the dendroview.

The pixel scaling determines how tall and wide the boxes are in both the zoom and global views. Basically, the larger the pixel scaling, the bigger the box. If the pixel scaling is less than one, the rows are averaged. This can make your data look better, since missing values disappear.

The contrast is the expression value which corresponds to fully induced. Any values greater than this will appear to be the induced color, and values between this and zero will appear to be a color between the zero and up color. The contrast is similarly used to color repressed boxes.

The color part allows you to set the up, down, zero and missing colors. You can double-click the boxes to get a color selection dialog, click a preset to load a color, and load and store color sets to files.

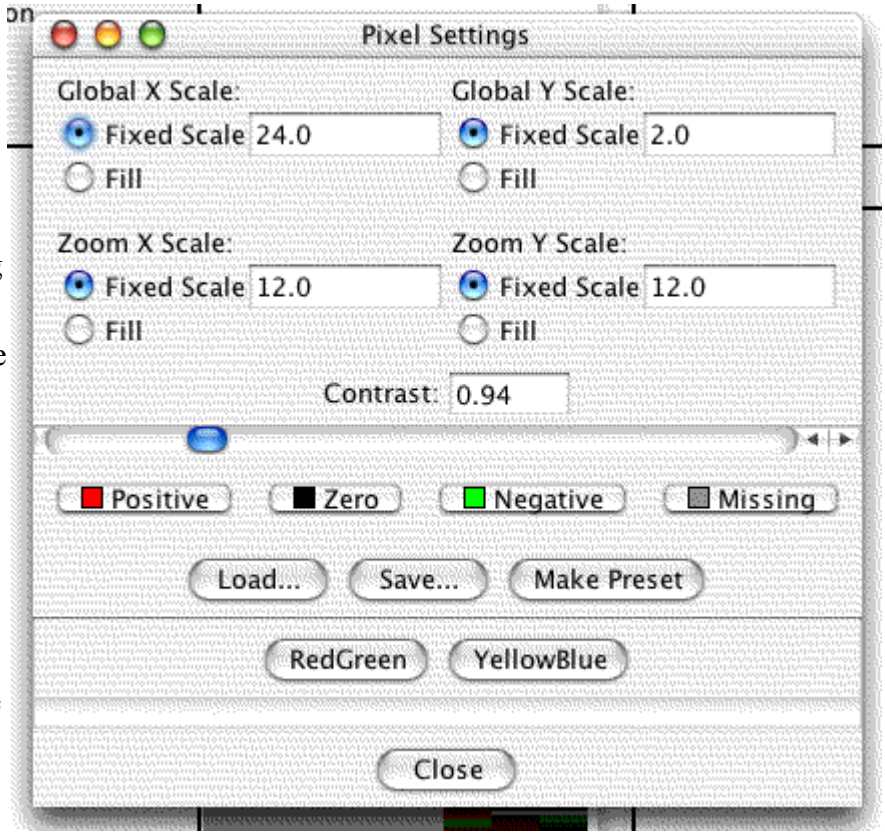

Figure F.2. Dendrogram Pixel Settings Dialog

Loading and saving of presets does not work because applets have no access to the hard drive for security.

Dendrogram

The Dendrogram is one of three views which can be created in the LinkedView application. It is also the bulk of the TreeView application. The dendrogram has a lot of components, so I've gone ahead and given them names, so that the description of the features will not be confusing.

The Dendrogram is the only view available via our applet.

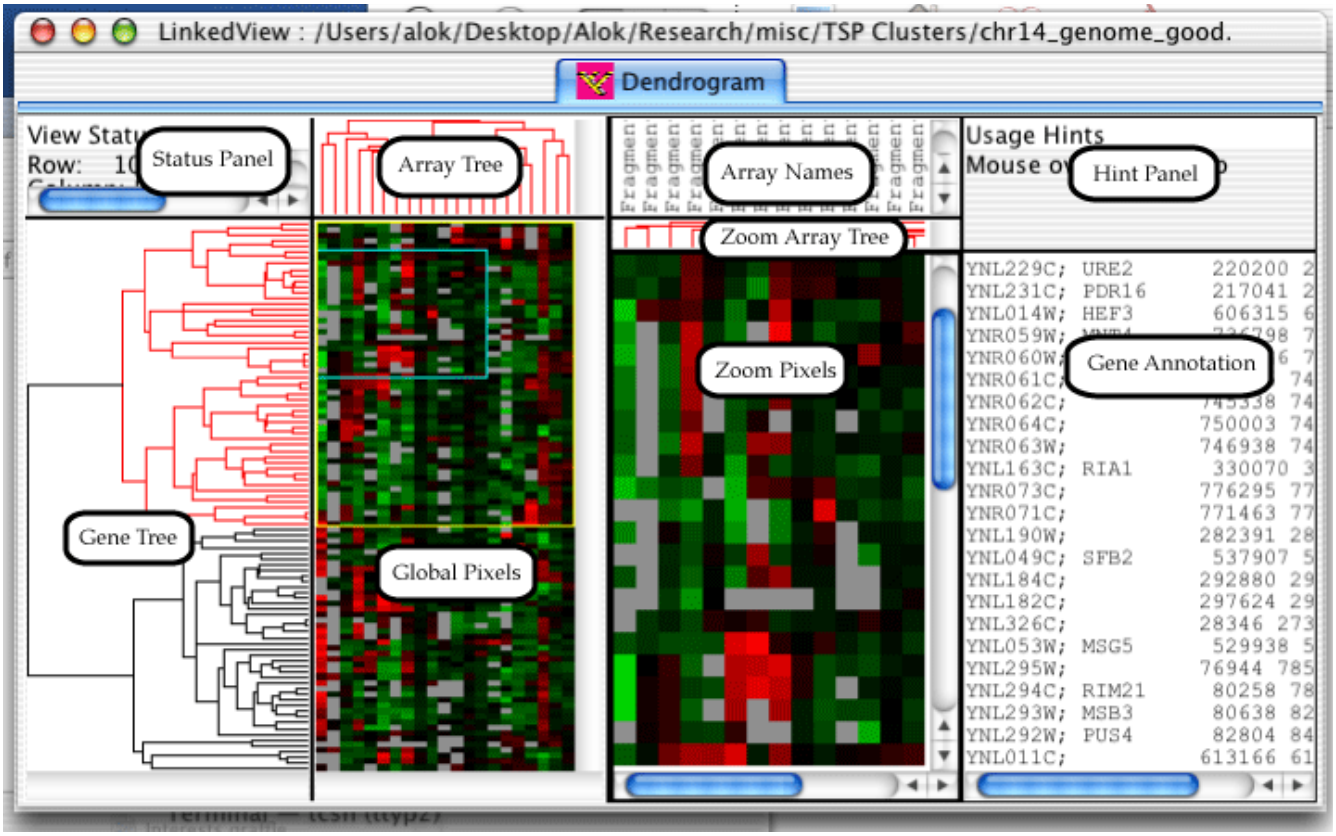

Figure 2.10. Dendrogram Component Layout

Informational Panels

The status panel displays different information depending upon which component the mouse is over. The hint panel displays hints on how to use the component.

Selection in Dendrogram

Genes can be selected in the dendrogram by either clicking and dragging on the global pixels or by clicking on a node in the gene or array trees. A zoomed in view of the selected genes will appear in the Zoom Pixels pane, a yellow rectangle will appear on the global view indicating which genes are selected, and a blue rectangle will appear inside the yellow rectangle indicating which genes are currently visible in the Zoom Pixels pane.

Holding down the shift key while dragging on the global pixels will cause the exact range of arrays to be selected; by default all arrays are selected. Once a range is selected, pressing the arrow keys moves the selected rectangle around. Holding down the control key while pressing the arrow keys will grow and shrink the selected rectangle.

Clicking on a node in the array or gene trees will select all descendants of the node. The selected node and descendants will be colored in red. At this point, pressing the arrow keys will change which genes

are selected. Up will select the parent of the current node, left and right will select the left and right children, and down will select the child with more descendants.

Url Linking in Dendrogram

Provided the url link settings (described in the section called “Url Settings”) are set appropriately, clicking on a gene annotation or an array name will cause a browser window to open with details on the gene or array.

References

Backstage with a command performer. Rockerfeller University News Release February 18, 2003. http://runews.rockefeller.edu/index.php?page=engine&id=103

Biology 2.0. The Economist Jun 17, 2010. http://www.economist.com/node/16349358?story_id=16349358

Collins FS, Green ED, Guttmacher AE, Guyer MS (2003). A vision for the future of genomics research. Nature 422: 835-847.

Davie JR and Spenser VA (1999). Control of histone modifications. Journal of Cellular Biochemistry 75: 141-148.

D'haeseleer P (2005). How does gene expression clustering work? Nature Biotechnology 23: 1499-1501.

De Hoon, MJL, Imoto S, Miyano S (2002). Statistical analysis of a small set of time-ordered gene expression data using linear splines. Bioinformatics 18**:** 1477–1485.

De Hoon MJL, Imoto S, Miyano S (2010). C Clustering Library. http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm

Ensembl Genome Browser. http://ensemble.org

Fischer A, et al. (2010). Targeting the correct HDAC(s) to treat cognitive disorders. Trends in Pharmacological Sciences 31: 605-617.

Galaxy-Central. http://getgalaxy.org

Galaxy Documentation. https://bitbucket.org/galaxy/galaxy-central/wiki/

Gentoo. http://gentoo.org

Guan J, et al. (2009). HDAC2 negatively regulates memory formation and synaptic plasticity. Nature 459: 55-63.

Holliday R (2006). Epigenetics: A Historical Overview. Epigenetics 1: 76-80.

Kazantsev AG and Thomson LM (2008). Theraputic application of histone deacetylace inhibitors for central nervous system disorders. Nature Reviews Drug Discovery 7: 854-868.

Mercurial revision control tool. http://mercurial.selenic.com

Microsoft Visual Studio Express. http://microsoft.com/express

Olins AL and Olins DE (1974). Spheroid Chromatin Units (v Bodies). Science 183: 330-332.

Oracle VM Virtualbox. http://virtualbox.org

Python. http://python.org/

Rogers JL and Nicewander WA (1988). Thirteen Ways to Look at the Correlation Coefficient. The American Statistician 42: 59-66.

Saldanha AJ (2004). Java Treeview—extensible visualization of microarray data. Bioinformatics 20: 3246-3248.

Sharma RP, Grayson DR, Gavin DP (2008). Histone deactylase 1 expression is increased in the prefrontal cortex of schizophrenia subjects: Analysis of the National Brain Databank microarray collection. Schizophrenia Research 98: 111-117.

Snedecor, GW and Cochran, W. G. (1989). Statistical Methods. Iowa State University Press.

Wilkinson L and Friendly M (2009). The History of the Cluster Heat Map. American Statistician 63: 179-184.

Yang Y, et al. (2006). Altered Levels of Acute Phase Proteins in the Plasma of Patients with Schizophrenia. Analytical Chemistry 78: 3571-3576.